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

# Psychometric Evaluation of Cloze Tests with the Rasch Model

Safa Mohammed Abdulridah Dhyaaldian[1], Sura Hasan Al-Zubaidi[2]*, Dhameer A. Mutlak[3], Nour Raheem Neamah[4], Ali A. Mohammed Ali Albeer[5], Doaa A. Hamad[6], Saad Fadhil Al Hasani[7], Mustafa Musa Jaber[8(a,b)], Hatem Ghaleb Maabreh[9]

Received: March 2022 Accepted: May 2022

**Abstract**
Cloze tests are gap-filling tests designed to measure overall language ability and reading comprehension in a second language. Due to their ease of construction and scoring, cloze tests are widely used in the context of second and foreign language testing. Previous research over the past decades has shown the reliability and validity of cloze tests in different contexts. However, due to the interdependent structure of cloze test items, item response theory models have not been applied to analyze cloze tests. In this research, we apply a method to circumvent the problem of local dependence for analyzing cloze tests with the Rasch model. Using this method, we applied the Rasch model to a cloze test composed of eight passages each containing 8-15 gaps. Findings showed that the Rasch model fits the data and thus it is possible to scale persons and cloze passages on an interval unidimensional scale. The test had high reliability and was well-targeted to the examinees. Implications of the study are discussed.

*Keywords*: cloze test, local item dependence, Rasch model, unidimensionality, validity

## 1. Introduction

A cloze test is a gap-filling test in which every $n^{th}$ word is deleted in a passage. Each deleted word is replaced with a blank or some dashes in the text. Examinees have to read the passage and fill in the missing words. This type of cloze test is called fixed ratio cloze test. In another type of cloze test, referred to as rational deletion cloze, specific vocabulary items or parts of speech are deleted. One point is allotted for every correct word supplied by an examinee. The total score on a cloze test is an indication of the examinee's overall language ability (Hughes,

[1] University of Warith Al-Anbiyaa, Karbala, Iraq
[2] Anesthesia Techniques Department, Al-Mustaqbal University College, Babylon, Iraq. Email: Sura.hasan.hasnawi@mustaqbal-college.edu.iq
[3] Al-Nisour University College/Baghdad/Iraq
[4] Al-Manara College for Medical Sciences (Maysan)/Iraq
[5] Department of Pharmacy, Al-Zahrawi University College, Karbala, Iraq
[6] Nursing Department, Hilla University College, Babylon, Iraq
[7] Al-Esraa University College, Baghdad, Iraq
[8] a. Department of Medical Instruments Engineering Techniques, Dijlah University College, Baghdad,10021, Iraq
b. Department of Medical Instruments Engineering Techniques, Al-Farahidi University, Baghdad,10021, Iraq
[9] People's Friendship University of Russia, Moscow, Russia

ISSN 2476-5880

2003) or reading comprehension (Alderson, 2000). The cloze procedure was invented by Taylor (1953) as a procedure for measuring text readability but it was later used as a measure of language competence. The word 'cloze' is the shortened form of the word 'closure'. In Gestalt psychology, closure is the process of perceiving incomplete things such as geometric figures (Oller, 1979). Taylor argued that the process of completing gapped texts is a kind of closure, hence, the cloze test.

## 2. Review of Literature

Over the past decades, cloze tests have been widely used in large-scale tests and in classroom assessments as measures of general language ability and reading comprehension (Abraham & Chapelle, 1992; Alderson, 1980; Kobayashi, 2002; Yamashita, 2003). Numerous studies and many researchers have also provided different kinds of validity evidence for the cloze tests. Correlational analyses have all demonstrated that cloze tests correlate with other tests of language ability including reading, writing, speaking, listening, vocabulary, and grammar (Oller, 1983). Criterion validity evidence has demonstrated substantial validity coefficients of .71 to .89 between cloze and standardized ESL (English as a Second Language) proficiency tests (Brown, 2013; Conrad, 1970; Darnell, 1970; Oller, 1972; Irvine et a1., 1974; Stubbs & Tucker, 1974). Recently, in a validation study of cloze test among Iraqi university learners of English as a foreign language (EFL), Sattar (2022) found relatively high correlation coefficients between a cloze test and tests of grammar ($r$=.70), vocabulary ($r$=.60), reading comprehension ($r$=.68), and the combined score of vocabulary plus grammar plus reading comprehension ($r$=.78). Zare and Boori (2018) found a strong correlation of .81 between cloze test and reading comprehension but Yazdinejad and Zeraatpishe (2019) found a moderate correlation of .48 between reading comprehension and cloze test. They stated that the low reliability of their tests was the reason for their relatively small correlation. When they corrected the correlation coefficient for attenuation, the correlation increased to .81. Reliability coefficients for cloze tests have also been very high in the magnitudes of .80 to.90 (Brown, 1983/2013).

Exploratory factor analysis studies have also shown that cloze tests load on a general language proficiency factor along with other language ability tests (Oller, 1983; Sattar, 2022). Other researchers have tried to find out if cloze tests measure language competence beyond the knowledge of sentence-level grammatical structures and tap into macro-level textual competence of the examinees. Ramanauskas (1972), for example, compared the performance of native English speakers on a cloze test in which the sentences were in original order and another version in which the sentences were randomly rearranged. Subjects significantly performed better in the cloze test in which the sentences were intact. Chihara, Oller, Weaver, and Chavez-Oller (1977) replicated Ramanauskas' study (1972) with native and nonnative speakers of English and found that both groups performed better on the intact passages. Oller (1975) showed that as the amount of context around gaps increased from 5 to 50 words, the average cloze item scores of native English speakers increased too. These results show that the cloze is sensitive to linguistic contexts longer than a sentence and thus cloze tests measure macro-level language abilities. In other words, cloze is not

just a test of micro-level grammar and vocabulary and taps into higher-order skills that operate beyond individual sentences. These findings are also evidence of the construct validity of cloze as they clearly show what is actually measured by a cloze test and what abilities underlie performance on cloze items.

Although different kinds of validity evidence have been accumulated for cloze tests, evidence based on fit to item response theory (IRT) models has not been provided so far[10]. The reason for this is that cloze items, i.e., gaps nested within passages, are interdependent and this is a violation of the local independence assumption of IRT models. IRT models have two basic assumptions, namely unidimensionality and local independence (Hambleton & Swaminathan, 1985). Unidimensionality states that all the items should measure a single latent trait while local independence states that after conditioning out the impact of the latent trait, the items should be uncorrelated. If the items remain correlated after factoring out the latent trait, it means that the test measures another dimension which is irrelevant to the target dimension we aim to measure. This is also a violation of unidimensionality and a piece of evidence against validity. Baghaei and Ravand (2016/2019) and Zhang (2010) are perhaps the only researchers who analyzed cloze tests with IRT models. They used testlet response theory (Bradlow et al., 1999), a bifactor multidimensional IRT model, to examine local item dependence (LID) in cloze tests. Their findings showed that cloze tests produce substantial levels of local dependence. Other than that, we could not find any other study in which cloze tests are analyzed with IRT models.

In this study, we analyzed a cloze test battery with the one-parameter logistic item response theory (1-PL) or the Rasch model (Rasch 1960/1980). Following the C-Test literature (Klein-Braley, 1985), we suggest developing a cloze test battery to solve the LID problem. A cloze test battery is a collection of 4-8 independent cloze passages each with 10 to 30 gaps. Theoretically and practically, there is no limitation on the number of gaps within each passage. By considering each passage as a super-item (Eckes & Baghaei, 2015) we can enter each passage into the IRT analysis as a polytomous item. Using this modeling strategy, the unit of analysis is passages instead of individual gaps and the LID problem does not arise. To our knowledge, this is the first study on the scalability of cloze tests with the Rasch model.

## 3. Method

### *3.1 Participants*

A sample of 178 students (97 female) studying English at Al-Nisour University College, Baghdad took the cloze tests. Their age ranged from 21 to 33 (M=23.69, SD=3.85). The cloze tests were administrated as a mid-term exam in a reading comprehension course in six parallel classes.

[10] As explained in the following sections, bifactor Rasch and IRT models, including testlet response theory, have been used to examine the magnitude of local dependence in cloze tests but unidimensional Rasch or IRT models have not been employed for evaluating the fit of cloze tests to such models and scale items and persons.

*3.2 Instrument*

A fixed ratio English cloze test battery containing eight independent passages was employed in this study. To construct the cloze tests, reading compression passages from the British Council website were used (https://learnenglish.britishcouncil.org/). The reading comprehension exercises on the British Council website are presented in five levels of A1, A2, B1, B2, and C1. For the purposes of this study, four passages were selected from the B1 level and four 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. Every 7$^{th}$ word was deleted in the passages which resulted in 8-15 gaps in each passage. The first and the last sentences remained intact to provide some contextual clues for text processing.

*3.3 Procedures*

To provide validity evidence for the cloze test battery, the Rasch measurement model (Rasch 1960/1980) was employed. As explained above, to solve the problem of local item dependence in cloze tests, each passage was considered a polytomous item with 9 to 16 response categories. The number of gaps in the passages varied between 8 and 15 but since zero is also a possible score, the number of response categories is one plus the maximum number of gaps in each passage. Masters' (1982) partial credit model which accommodates polytomous items with a different number of response categories was used to analyze the data. 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. Item difficulty estimates ranged from -.42 to .44 logits. 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 good fit to the Rasch model. This is evidence that the items contribute to the definition of the latent trait that we aimed to measure, i.e. reading comprehension. Put differently, the items conform to the logistic functional shape assumed by the Rasch model for the behavior of the items.

**Table 1**

*Item measures and fit statistics for the eight cloze test passages*

| Item | Difficulty | S.E. | Infit MNSQ | Outfit MNSQ | Pt. Meas. Cor. |
|---|---|---|---|---|---|
| 1 | -.06 | .04 | 1.19 | 1.14 | .72 |
| 2 | .15 | .05 | .82 | .78 | .76 |
| 3 | .44 | .06 | .98 | .96 | .64 |
| 4 | .07 | .04 | 1.03 | .99 | .79 |
| 5 | -.16 | .04 | .92 | .95 | .79 |
| 6 | -.42 | .04 | .91 | .91 | .79 |
| 7 | -.14 | .04 | .99 | .99 | .76 |
| 8 | .41 | .05 | 1.23 | 1.25 | .64 |

Note: Pt. Meas. Cor. =Point-Measure Correlation; MNSQ= Mean Square

ISSN 2476-5880

Figure 1 shows the item characteristic curve (ICC) for Item 7 which is the best fitting item with outfit mean square value of .99. As the Figure shows, the empirical ICC (the blue line) is very close to the logistic S-shaped theoretical ICC (the red curve) imposed by the model. Therefore, the Item 7 fits the Rasch model's predictions. The lines on the two sides of the ICC show the upper and lower 95% confidence intervals. If the ICC is within these two lines, it is an indication that the items fit the Rasch model.

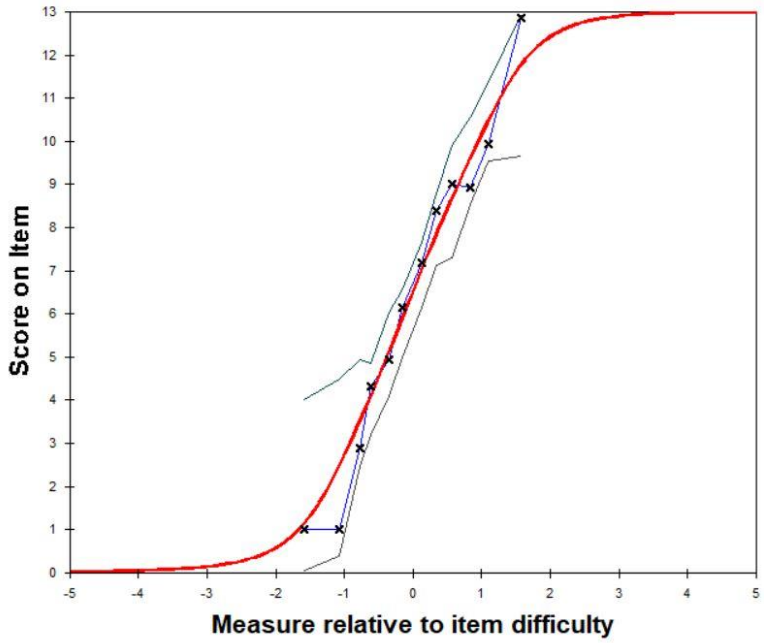


*Figure 1.*
Empirical and theoretical ICCs for cloze Item 7

Figure 2 is the ICC for Item 8 which is the worst fitting item with an outfit mean square value of 1.25. Figure 2 shows that the empirical ICC deviates at some points from the model predicted S-shaped curve. Nevertheless, the deviations are still within the acceptable boundaries, i.e., within the 95% confidence intervals. Therefore, it is possible to estimate person and item parameters on a unidimensional interval scale.

ISSN 2476-5880

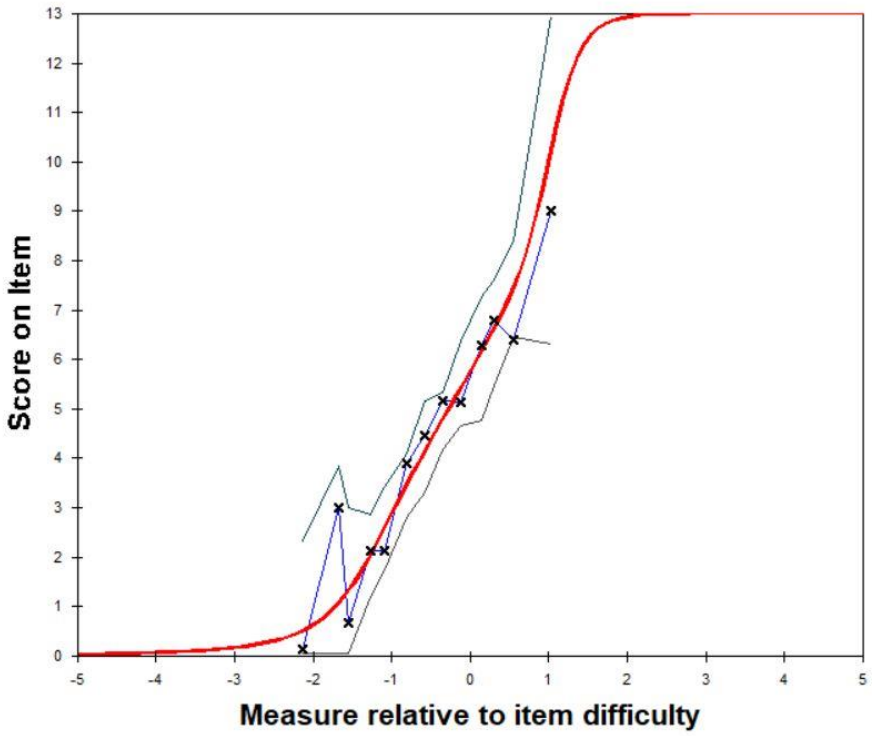


*Figure 2*
Empirical and theoretical ICCs for cloze Item 8

Figure 3 shows the bubble chart for the eight cloze super-items (passages). Each circle represents an item. Items on the top are harder and items towards the lower end of the scale are easier items. The *x*-axis indicates the outfit mean-sqaure. Perfect outfit is equal to 1. Items close to the horizontal line, which represents 1, are better fitting items and items further from the line have worse fit. The bubble chart shows that Item 3 is the hardest item and Item 6 is the easiest item. Item 7 is exactly on the horizontal line in the middle and thus has the best fit and Item 8 is the farthest item from the horizontal line and, therefore, has the worst fit (based on outfit mean-square).

ISSN 2476-5880

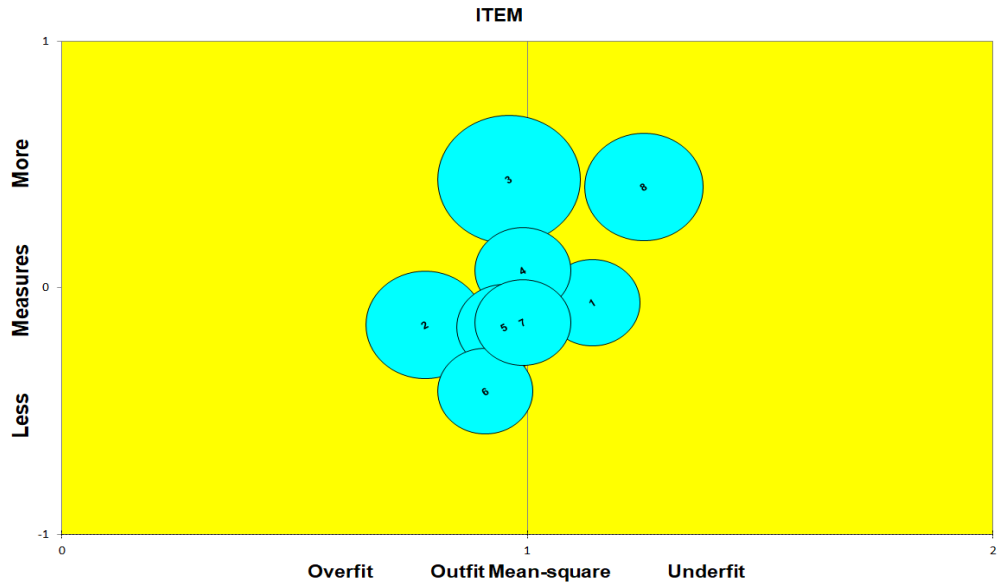


*Figure 3*
Bubble chart representing item measures and outfits

S.E. shows the standard error of estimation for the item difficulty parameters. The lower the S.E.s, the more precise the estimations are. As Table 1 shows, the standard errors are very small (.04 to .06) which indicate that the item difficulties are estimated very precisely. Further analyses showed a Rasch separation reliability of .88 for the test. Point-measure correlations are the correlations between the items and person ability measures. Higher correlations indicate higher discrimination for the items.

Figure 4 is the Person-Item map or the Wright map (Wilson, 2005) for the data. The vertical line shows the logit scale. The '#' and dots on the left represent persons (ach '#' is 2 persons and each dot is one person). The letters CLZ represent cloze items. Since there are eight items, we have CLZ 1 to CLZ 8 on the map. As the items are polytomous, the location of the thresholds for the items are also shown. For instance, CLZ1.12 shows the cloze item 1, threshold 12. Because the items are polytomous, showing the location of the items alone does not help to figure out test targeting. The location of the thresholds indicate the operational range of the scale (Baghaei, 2014). Figure 4 shows that the thresholds cover a wide range of the construct which is wider than the distribution of persons. The distribution of the persons matches exactly the distribution of item thresholds. Therefore, we can conclude that the test is well-targeted to the examinees and provides accurate measures of their reading ability.

ISSN 2476-5880

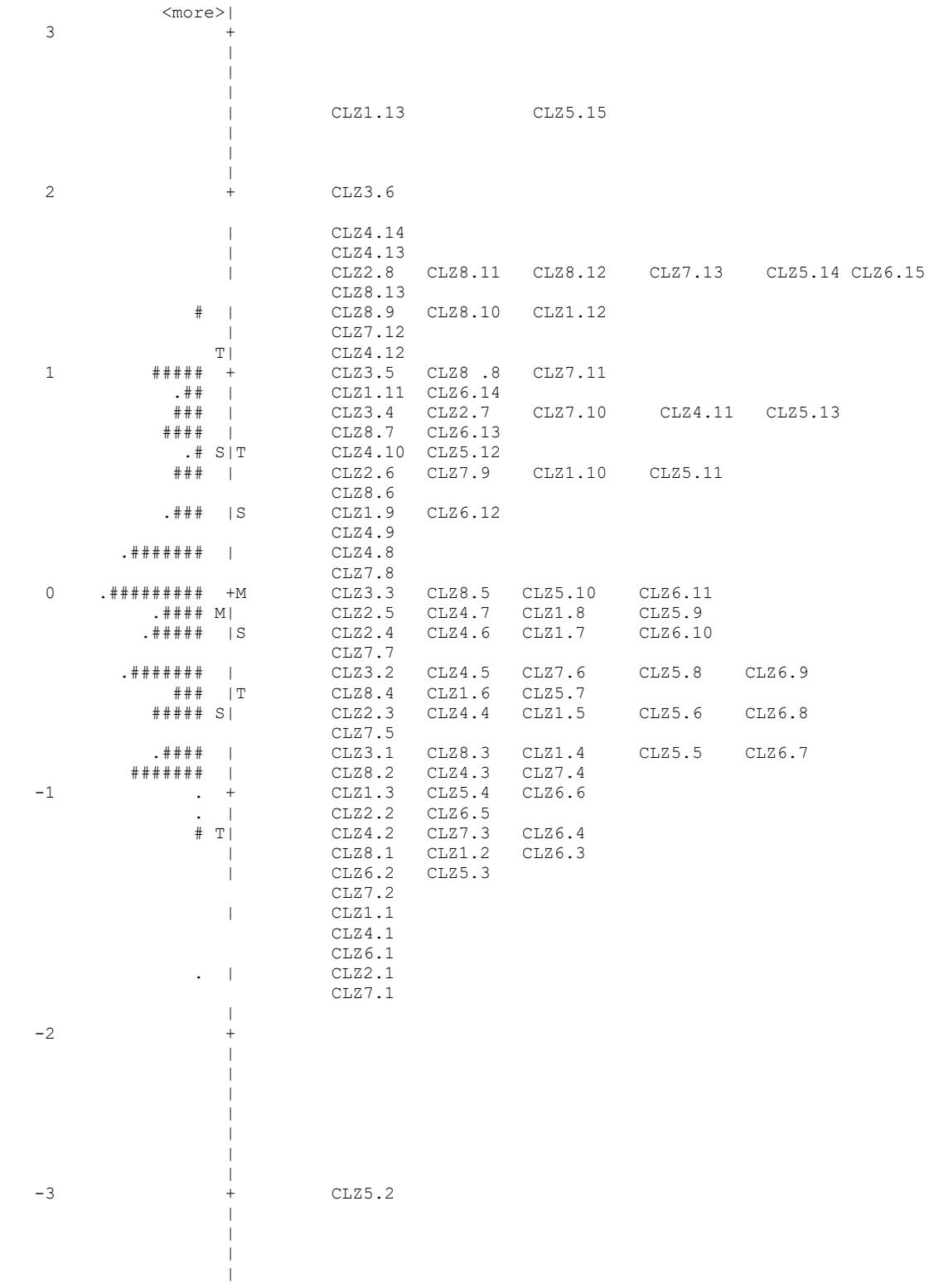


*Figure 4*
Person-Item map with Rasch-Thurstone thresholds

## 5. Conclusion

The purpose of this study was to evaluate the fit of a cloze test to the unidimensional Rasch model. The Rasch model is frequently used for the analysis and validation of educational and psychological tests and for scaling respondents. This is because the Rasch model enjoys appealing properties including parameter separability and existence of a sufficient statistic to estimate person and item parameters independently of each other (Baghaei et al. 2017; Fischer, 2006). These features are unique to the Rasch model and are not shared with the other IRT models such as the 2-parameter logistic model and the 3-parameter logistic model.

Cloze tests are commonly used in English as a foreign language contexts in classroom testing and in large and medium scale assessments. However, to the best of our knowledge, the fit of cloze tests to IRT models has not been investigated before. The reason for this gap in the cloze test literature is that the cloze items, i.e., the gaps are locally dependent and this is a violation of the local independence assumption of IRT models. Farhady (1983) even argued that due to the interdependence of cloze items, the application of internal consistency reliability coefficients like Cronbach'a alpha or KR20 is problematic for cloze items.

To solve the LID problem in cloze tests, we followed the strategy employed by C-Test researchers (Forthmann et al., 2020). A C-Test is a variation of the cloze test in which half of the words are deleted instead of the whole words and the rate of deletion is two. That is, the second half of every other word is deleted. A C-Test is composed of 4-8 independent passages. To solve the LID problem in C-Test, researchers suggested considering each passage as a polytomous item or a super-item and entering passages into the analysis (Raatz, 1984). By using this strategy, the problem of LID is solved as the unit of analysis is passages and not the gaps and the passages are independent.

To make cloze tests analysable with the Rasch model, we construed a cloze test battery using eight different and independent cloze passages. Each passage was then entered into the Rasch analysis as a super-item with 9 to 16 response categories. Masters (1982) Rasch partial credit model was applied to analyze the data. Findings showed that the Rasch model fits cloze test passages. This is an indication that the data are unidimensional and it is therefore possible to locate items and persons on an interval scale with a cloze test. Furthermore, the fit of data to the Rasch model shows that item and person raw scores can be used to place items and examinees on an ordinal scale. Examination of the Item-Person map or the Wright map showed that thresholds of the cloze super-items cover a wide range of the ability scale and thus all examinees at varying levels of ability can be measured with precision. By considering each passage as a super-item not only cloze tests can be analyzed with IRT models but also the application of internal consistency reliability coefficients, if applied to passages, is permissible.

In this study, for the first time, we suggested and applied a strategy to scale cloze tests with the Rasch model. Our findings showed that this strategy works and solves the problem of LID in cloze tests. We also used graphical evaluation of item fit which is conducted by comparing empirical and theoretical ICCs. While checking ICCs is very common in nonparametric IRT models (Baghaei, 2021; Firoozi, 2021; Tabatabaee-Yazdi et al., 2021), it is rarely used in

parametric IRT models such as the Rasch model. Checking the ICCs shows whether the items are monotonic and if they conform to the logistic shape imposed by the model. Future research should examine the fit of cloze tests to other types of Rasch models including the Rasch Poisson counts model for speeded tests (Baghaei & Doebler, 2019), linear logistic test model (Fischer, 1973; Hohensinn & Baghaei, 2017) and multidimensional Rasch model (Adams et al., 1997; Baghaei, 2013).

**References**

Abraham, R. G., & Chapelle, C. A. (1992). The meaning of cloze test scores: An item difficulty perspective. *Modern Language Journal, 76*, 468–479.

Adams, R. J., Wilson, M. R., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1–23.

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

Alderson, J. C. (1980). Native and non-native speaker performance on cloze tests. *Language Learning, 30*, 59–76.

Baghaei, P. (2021). *Mokken scale analysis in language assessment.* Münster: Waxmann Verlag.

Baghaei, P., & Ravand, H. (2019). Method bias in cloze tests as reading comprehension measures. *SAGE Open*. https://doi.org/10.1177/2158244019832706

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

Baghaei, P. (2014). Development and validation of a C-Test in Persian. In R. Grotjahn (Ed.). *Der C-Test: Aktuelle Tendenzen/The C-Test: Current trends* (pp.299-312). Frankfurt/M.: Lang.

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.

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

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

Brown, J. D. (1983). A closer look at cloze: Validity and reliability. In Oller, J. W. Jr. (Ed.), *Issues in language testing* (pp. 237-250). Rowley, MA: Newbury House.

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.

Conrad, C. A. (1970). *The Cloze Procedure as a Measure of English Proficiency*. Unpublished M.A thesis, University of California, Los Angeles.

Darnel, D. K. (1970). Clozentropy: A procedure for testing English language proficiency of foreign students. *Speech Monographs, 37*, 36-46.

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

Farhady, H. (1983). New directions for ESL proficiency testing. In J. W. Oller, Jr. (Ed.), *Issues in language testing research* (pp. 253-269). Rowley, MA: Newbury House.

Firoozi, F. (2021). Mokken Scale Analysis of the reading comprehension section of the International English Language Testing System (IELTS). *International Journal of Language Testing, 11*(2), 91-108.

Fischer, G. H. (2006). Rasch models. In Rao, C. & Sinharay, S. (Eds.). *Handbook of statistics, Volume 26: Psychometrics* (pp. 979-1027). Amsterdam, The Netherlands: Elsevier.

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

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

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

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

Hughes, A. (2003). *Testing for language teachers* (2nd Ed.). Cambridge: Cambridge University Press.

Irvine, P., Atai, P., & Oller, J.W. Jr. (1974). Cloze, dictation, and the Test of English as a Foreign Language. *Language Learning, 24*, 245-252.

Klein-Braley, C. (1985). A cloze-up on the C-test: A study in the construct validation of authentic tests. *Language Testing, 2*, 76–104.

Kobayashi, M. (2002). Method effects on reading comprehension test performance: Text organization and response format. *Language Testing, 19*, 193–220.

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*(2), 149-174.

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). Rowley, MA: Newbury House.

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). Washington, D.C.: TESOL.

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

Oller, J. W., Jr. (1972). Scoring methods and difficulty levels for tests of proficiency in English as a second language. *Modem Language Journal, 56*, 151-158.

Ramanauskas, Si. (1972). The responsiveness of cloze readability measures to linguistic variables operating over segments of text longer than a sentence. *Reading Research Quarterly, 8*(1), 72-91.

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.

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

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.

Stubbs, J. B., & Tucker, G. R. (1974). The cloze test as a measure of English proficiency. *Modern Language Journal, 58*, 239-242.

Tabatabaee-Yazdi, M., Motallebzadeh, K., & Baghaei, P. (2021). A Mokken Scale Analysis of an English reading comprehension test. *International Journal of Language Testing, 11*, 132–143.

Taylor, W. L. (1953). Cloze procedure: a new tool for measuring readability. *Journalism Quarterly, 30*, 415-453.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum.

Yamashita, J. (2003). Processes of taking a gap-filling test: Comparison of skilled and less skilled EFL readers. *Language Testing*, *20*(3), 267–293.

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.

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