Tabaran Institute of Higher Education
International Journal of Language Testing
Special Issue, 2023: *Advanced Psychometric Methods in Language Testing*
Research Note

ISSN 2476-5880

Psychometric Modelling of Reading Aloud with the Rasch Model

Kuanysh Syman[1], Hajir Mahmood Ibrahim Alallo[2], Aisha Mohammed[3], Aalaa Yaseen Hassan[4], Osama Wael Suleiman[5], Yusra Mohammed Ali[6], Maabreh Hatem Ghaleb[7], Kupriianova Anastasiia Mikhailovna[7], Emaimo Alice John[7], Goncharova Velichka Georgievna[7]

Abstract
Reading aloud is recommended as a simple technique to measure speaking ability (Hughes & Hughes, 2020; Madsen, 1983). Reading aloud is currently used in the Pearson Test of English and a couple of other international English as a second language proficiency tests. Due to the simplicity of the technique, it can be used in conjunction with other techniques to measure foreign and second language learners' speaking ability. One issue in reading aloud as a testing technique is its psychometric modeling. Because of the peculiar structure of reading-aloud tasks, analyzing them with item response theory models is not straightforward. In this study, the Rasch partial credit model (PCM) is suggested and used to score examinees' reading-aloud scores. The performances of 196 foreign language learners on five reading-aloud passages were analyzed with the PCM. Findings showed that the data fit the RPCM well and the scores are highly reliable. Implications of the study for psychometric evaluation of reading aloud or oral reading fluency are discussed.

Keywords: Rasch partial model; reading aloud; speaking test; validation

1. Introduction

Reading aloud is a simple controlled technique which may be used along with other techniques to assess foreign and second language learners' oral abilities (Brown, 2018; Harris, 1968; Hughes & Hughes, 2020; Madsen, 1983; Underhill, 1987). The technique involves asking examinees to read short passages or even sentences loud to the examiner. A number of international tests including Versant®, the Test of Spoken English (TSE®), and the Pearson Test of English have included reading aloud as a measure of oral abilities. Research by Versant developers has shown that reading aloud has very high correlation coefficients with traditional interview-based tests of

[1] Institute of Natural Sciences and Geography, Abai Kazakh National Pedagogical University, Almaty, Kazakhstan
[2] English Department, Ahl Al Bayt University, Kerbala, Iraq
[3] College of Education, Al-Farahidi University, Baghdad, Iraq
[4] Al-Nisour University College, Baghdad, Iraq
[5] English Department, AlNoor University College, Nineveh, Iraq
[6] Department of Medical Laboratory Technics, Al-Zahrawi University College, Karbala, Iraq
[7] Peoples Friendship University of Russia, Moscow, Russia

oral abilities (around .75) (Balogh & Bernstein, 2007; Bernstein et al., 2010; Cascallar & Bernstein, 2000; Downey et al., 2008).

The advantage of the reading-aloud technique is that it is easily administrated. Examinees have to read either a connected short passage or several independent sentences. The test is uniform and standardized for all the examinees because they all read the same material which leads to high reliability (Brown, 2018). The scoring is also simple and quick. The disadvantages are that: the technique is inauthentic. We rarely read aloud in real life. Besides, it very much depends on reading skill and cannot be used for children who have not yet learned to read or with illiterate people. Next, even educated native speakers differ in their ability to read texts aloud. And finally, the technique only allows the assessment of micro skills of pronunciation, intonation, and sentence stress patterns and does not measure interaction and appropriate responses (Madsen, 1983; Underhill, 1987).

The scoring could focus on pronunciation and fluency by rating these two features on a Likert scale as explained in the *Manual of American English Pronunciation* (Prator, 1972). In the other method of scoring, a carefully selected number of words or phrases are only scored within a connected passage. That is, a passage is selected and a number of words or phrases in the passage are marked as test items. These may include technical words, idiomatic expressions, contractions, liaisons, minimal pairs, and words or sounds that are known to be difficult to produce orally. The examinee reads the passage and the assessor only marks and records the correct production of these preselected items and the rest of the passage is not considered in the scoring.

Due to the peculiar structure of reading aloud, psychometric modeling of this testing technique with the item response theory (IRT) models is not straightforward. Interestingly, one of the first applications of the Rasch model by Georg Rasch was to oral reading fluency data of Danish students (Baghaei & Doebler, 2019). Rasch (1960/1980) employed the Rasch Poisson Counts Model but the application of this model is limited because of the unavailability of user-friendly software and other technical issues (see Baghaei and Doebler, 2019 for details). In this study, we aim to apply the partial credit model (PCM, Masters, 1982) to reading aloud data.

2. Method

2.1. Participants

The participants of the study were 196 (115 female) undergraduate students of philology at the Abai Kazakh National Pedagogical University, Almaty, Kazakhstan. Their age range was 18 to 41 (M=21.54, sd =4.98).

2.2 Instruments

Five independent passages from a popular reading comprehension book for second-language learners were selected. The passages were about bicycles (96 words), medicine (112 words), narcissism (129 words), fast food (151 words), and autism (139 words).

2.3 Procedure

The passages were first handed out to the examinees on a sheet of paper. They were given 10 minutes to read the passages silently to get familiar with the meaning and prepare the sections of text which needs attention, as recommended by Underhill (1987). Then the sheets were collected. Afterward, examinees entered an examination room one by one where two assessors asked them to read the passages aloud. The number of errors for each passage was counted independently by both assessors. For scoring, all the words in a passage were considered an item. That is, all the errors of pronunciation and stress were counted, irrespective of word type. The correlations between the numbers of errors counted by the two assessors were computed for the four passages separately. The Pearson coefficients of correlation were .96, .94, .97, .96, and .95, for Text 1, Text 2, Text 3, Text 4, and Text 5, respectively. These values indicate a high level of agreement between the two assessors in counting the number of errors.

3. Results and Discussion

The Rasch partial credit model (PCM) of Masters (1982) was employed to analyze the data. Since, the number of words in each passage is different, here we modeled the number of errors in each passage. As the number of errors is an undefined value, the exact number of categories in each item or passage is not known. Therefore, the PCM, which does not make any assumptions as regards the number of categories in each item, was used. Recently, Hussein et al. (2022) Dhyaaldian, Hassan, et al. (2022), Dhyaaldian, Kadhim, et al. (2022), and Effatpanah and Baghaei (2022) used the PCM to analyze foreign language dictation, C-Test, cloze test data, and cloze elide, respectively. All these test types have a similar structure to reading aloud data. Winsteps Rasch computer program (Linacre, 2022a) was used to estimate the PCM.

Table 1 shows the item (passage) difficulty estimates, their standard errors of measurement, infit and outfit mean square values, and their point-measure correlation coefficients. According to Table 1, the passages differ in difficulty. Note that here the errors were modeled. Thus, Item 3 which has the highest measure, in fact, has the smallest count of errors and is the easiest passage. Passage 1 which has the highest count of errors is the hardest passage to read.

The fit indices indicate that all five passages fit the unidimensional Rasch measurement model. This indicates that all the items work together to define a latent variable of oral reading ability and examinees can be located on this line on an interval scale (Wright & Stone, 1979). Point-measure correlations show the correlation between the item and the person Rasch parameter estimates. The higher the correlation, the more related the item is to the test. This is equivalent to classical item-total correlation or item discrimination.

Table 1.
Item Measures and Fit Statistics for the Five Reading-aloud Passages

Item	Total Score	Diff.	SE	Infit MNSQ	Outfit MNSQ	Pt. Meas. Cor.
1	2414	-.24	.02	1.05	1.07	.85
2	1741	-.02	.02	1.14	1.08	.83
3	1001	.64	.02	.85	.72	.81
4	1934	-.16	.02	1.02	.97	.84
5	2019	-.23	.02	1.11	.92	.85

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

Table 2 shows the test and sample statistics. Table 2 indicates that the five reading-aloud texts have a very high reliability (*r*=.90) which is an indication that the candidates have been measured with very high precision. The person separation value of 2.92 means that the reading-aloud test battery consisting of five passages as a whole can identify three levels of reading-aloud ability strata. Item separation value of 12.58 shows that respondents have identified more than 12 levels of difficulty strata in the items. The high value for item separation is a sign that the reading-aloud passages vary in difficulty. Principal components analysis (PCA) of standardized residuals indicated that the eigenvalue in the first contrast is 1.6. This is a method of detecting unidimensionality. The strength of the component is smaller than 2 items which shows that the test is unidimensional (Baghaei & Cassady, 2014; Linacre, 2022b).

Table 2.
Overall Test and Sample Statistics

Reliability	.90
Person separation	2.92
Item separation	12.58
Mean (SD)	-.24 (.80)
Range	4.90
Eigenvalue in 1st Contrast	1.60

4. Conclusion

Reading aloud is a technique that may be used for measuring speaking ability. It is easy to administer and score and correlates highly with traditional measures of speaking that are hard and expensive to administer and score. Nevertheless, psychometric analysis of reading-aloud tasks with IRT models is not easy. This is due to the peculiar structure of reading-aloud tasks which do not lend themselves to psychometric analysis. Modern psychometric theory, namely, item response theory is designed to work at the level of individual items but reading aloud tasks are holistic and indivisible to independent items.

To solve the problem mentioned above, the use of several reading-aloud tasks was suggested. Each task or reading-aloud passage was considered a polytomous or a super-item with the number of errors in each passage as the score. Then the unit of analysis for IRT became passage with the number of errors in the passages as item scores to be fed to the IRT analysis. Previous research on such items has also shown that modeling the number of errors leads to better psychometric results (Baghaei et al., 2019; Nadri et al., 2019).

In this study, five independent reading-aloud passages were evaluated with the PCM (Masters, 1982) which is a polytomous Rasch model. Our findings revealed that reading aloud passages fit the unidimensional Rasch model and have very high reliability. This is the first study which demonstrates the psychometric quality of reading aloud with the Rasch model. Future studies may look into the psychometric quality of reading-aloud tasks with other IRT models The item-based scoring of reading-aloud tasks by scoring only specific selected items can open the way for the application of some IRT models which allows the examination of learners' strengths and weaknesses –by applying cognitive classification models (see Alallo et al., 2023; Boori et al., 2023; Effatpanah et al., 2023)– and linear logistic test model to identify sources of difficulty in reading aloud tasks (Baghaei & Kubinger, 2015; Effatpanah & Baghaei, 2021; Hohensinn & Baghaei, 2017; Fischer, 1973). The Rasch Poisson Counts Model (Rasch 1960/1980) is particularly apt for the analysis of reading aloud data.

References

Alallo, H. M. I., Mohammed, A., Hamid, Z. K., Hassan, A. Y., & Kadhim, Q. K. (2023). Examining attribute relationship using diagnostic classification models: A mini review. *International Journal of Language Testing*, (Special Issue: Advanced Psychometric Methods in Language Testing), 21-30. doi: 10.22034/IJLT.2022.368468.1212

Baghaei, P., & Cassady, J. (2014). Validation of the Persian translation of the Cognitive Test Anxiety Scale. *Sage Open, 4*, 1-11. doi: 10.1177/2158244014555113

Baghaei, P., & Doebler, P. (2019). Introduction to the Rasch Poisson Counts Model: An R tutorial. *Psychological Reports*, *122*(5), 1967-1994. doi: 10.1177/0033294118797577

Baghaei, P. & Kubinger, K. D. (2015). Linear logistic test modeling with R. *Practical Assessment, Research & Evaluation, 20,* 1-11.

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

Balogh, J., & Bernstein, J. (2007). Workable models of standard performance in English and Spanish. In Y. Matsumoto, D. Oshima, O. Robinson, & P. Sells (Eds.), *Diversity in language: Perspectives & implications* (pp. 217-229). Stanford, CA: CSLI Publications.

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing, 27,* 355-377. doi: 10.1177/0265532210364404

Boori, A. A., Ghazanfari, M., Ghonsooly, B., & Baghaei, P. (2023). The construction and validation of a Q-matrix for cognitive diagnostic analysis: The case of the reading comprehension section of the IAUEPT. *International Journal of Language Testing,* (Special Issue: Advanced Psychometric Methods in Language Testing), 31-53. doi: 10.22034/IJLT.2023.383112.1227

Brown, H. D., & Abeywickrama, P. (2018). *Language assessment: Principles and classroom practices* (3rd Ed.). New York: Pearson.

Cascallar, E., & Bernstein, J. (2000, March). *The assessment of second language learning as a function of native language difficulty measured by an automated spoken English test*. Paper presented at the American Association of Applied Linguistics Conference, Vancouver, BC, Canada.

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

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*, 107-117. doi: 10.22034/IJLT.2022.157128

Downey, R., Farhady, H., Present-Thomas, R., Suzuki, M., & Van Moere, A. (2008). Evaluation of the usefulness of the Versant for English test: A response. *Language Assessment Quarterly, 5*, 160-167. doi: 10.1080/15434300801934744

Effatpanah, F., & Baghaei, P. (2022). Evaluating different scoring methods for the speeded cloze-elide test: The application of the Rasch partial credit model. *The Quantitative Methods for Psychology, 18*(3), 241–254. doi:10.20982/tqmp.18.3.p41

Effatpanah, F. & Baghaei, P. (2021). Cognitive components of writing in a second language: An Analysis with the Linear Logistic Test Mode. *Psychological Test and Assessment Modeling, 63*, 13-44.

Effatpanah, F., Baghaei, P., & Boori, A. (2019). Diagnosing EFL learners’ writing ability: A diagnostic classification modeling analysis. *Language Testing in Asia, 9*(12). doi: 10.1186/s40468-019-0090-y

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359-374. doi: 10.1016/0001-6918(73)90003-6

Harris, D. P. (1968). *Testing English as a second language*. New York: McGraw-Hill.

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

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

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*, 118–127. doi: 10.22034/IJLT.2022.157129

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.

Madsen, H. (1983). *Techniques in testing*. New York, NY: Oxford University Press.

Nadri, M., Baghaei, P., & Zohoorian, Z. (2019). Analysis of the Ruff 2 & 7 Test of Attention with the Rasch Poisson Counts Model. *The Open Psychology Journal*, *12*, 7-11. doi: 10.2174/1874350101912010007

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

Prator, C. H. (1972). *Manual of American English pronunciation*. New York, NY: Holt, Rinehart & Winston.

Underhill, N. (1987). *Testing spoken language: A handbook of oral testing techniques*. Cambridge, England: Cambridge University Press.

Wright, B. D. & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.