Tabaran Institute of Higher Education ISSN 2476-5880
International Journal of Language Testing
Special Issue, 2023: *Advanced Psychometric Methods in Language Testing*

Multidimensional IRT Analysis of Reading Comprehension in English as a Foreign Language

Galiya Tatarinova[1]*, Nour Raheem Neamah[2], Aisha Mohammed[3], Aalaa Yaseen Hassan[4], Ali Abdulridha Obaid[5], Ismail Abdulwahhab Ismail[6], Hatem Ghaleb Maabreh[7], Al Khateeb Nashaat Sultan Afif[7], Shvedova Irina Viktorovna[7]

Abstract

Unidimensionality is an important assumption of measurement but it is violated very often. Most of the time, tests are deliberately constructed to be multidimensional to cover all aspects of the intended construct. In such situations, the application of unidimensional item response theory (IRT) models is not justifieddue to poor model fit and misleading results. Multidimensional IRT (MIRT) models can handle several dimensions simultaneously and yield person ability parameters on several dimensions which is helpful for diagnostic purposes too. Furthermore, MIRT models use the correlation between the dimensions to enhance the precision of the measurement. In this study, a reading comprehension test is modeled with the multidimensional Rasch model. The findings showed that a correlated 2-dimensional model has the best fit to the data. The bifactor model revealed some interesting information about the structure of reading comprehension and the reading curriculum. Implications of the study for the testing and teaching of reading comprehension are discussed.

Keywords: Bifactor model, Multidimensional IRT; Reading comprehension; Unidimensional IRT

1. Introduction

Theoretically, measurement in the physical and social sciences has been unidimensional (Bond et al., 2021), i.e., one construct is measured at a time. Conflating measures is not acceptable in measurement. No one combines measures of weight and height; each of these variables is

[1] Institute of Natural Sciences and Geography, Abai Kazakh National Pedagogical University, Almaty, Kazakhstan, 13, Dostyk Av., 050010 Almaty, the Republic of Kazakhstan; Orcid: 0000-0002-2520-0338, Email: tat_galiya@mail.ru
[2]Al-Manara College for Medical Sciences, Maysan, Iraq
[3]College of Pharmacy / Al-Farahidi University, Baghdad , Iraq
[4]Al-Nisour University College, Baghdad, Iraq
[5]Mazaya University College, Iraq
[6]English Department, Al-Noor University College, Nineveh, Iraq
[7]People's Friendship University of Russia, Moscow, Russia

reported separately (Bond et al., 2021). However, educational and psychological constructs could be very complex. Such constructs may be composed of several uncorrelated dimensions. Treating such constructs as unidimensional is an oversimplification of the issue and may lead to loss of information about the various aspects of the construct.

When complex educational and psychological data are modeled, squeezing all the items into a single unidimensional scale results in inaccurate parameter estimates and misleading substantive interpretations (Baghaei, 2012). Under such situations, the easiest and most straightforward modeling strategy is to analyze each subscale separately with a unidimensional measurement model which is referred to as the consecutive approach (Adams et al., 1997). The advantage of this model is that it is simple and recognizes the fact that the construct is composed of several distinct sub-constructs. It also provides separate person parameters for each of the sub-constructs which is useful for diagnostic purposes and individualized learning and decisions (Baghaei, 2012). The disadvantage is that since the number of items within each subscale is small, it leads to imprecise person parameter estimates on each subscale (Adams et al., 1997).

Multidimensional item response theory (MIRT; Reckase, 2009) models are useful psychometric measurement models for modeling complex constructs that are composed of several sub-constructs. These measurement frameworks model several dimensions simultaneously. The advantages of these models are that (1) there is no need to analyze each subscale separately with a unidimensional model one at a time, (2) these models are compensatory, i.e., information can be borrowed from one dimension to boost the precision of the estimates from another dimension, provided that they are correlated (Adams et al., 1997). In other words, with fewer items, we can have higher accuracy in measurement, and (3) MIRT models also help in examining the internal structure of latent constructs (Adams et al., 1997). As confirmatory measurement models, MIRT models allow the comparison of several structures (unidimensional, multidimensional, bifactor, higher order) for latent variables. The best fitting model can be selected based on the fit values offered by the models. And (4) MIRT models can be used to examine testlet and method effects (see Baghaei & Aryadoust, 2015; Baghaei & Ravand, 2016; Baghaei & Ravand, 2019).

The purpose of the present study is to evaluate the factorial structure of reading comprehension in English as a foreign language using the multidimensional Rasch model of Adams et al. (1997). To this end, a reading comprehension test was examined with three models: a unidimensional model, a 2-dimensional correlated model, and a 2-dimensional uncorrelated model.

2. Method

2.1. Participants

Participants of the study were 356 undergraduate philology students at the Abai Kazakh National Pedagogical University, Almaty, Kazakhstan. A total of 238 students were female and the rest were male. The participants were from two different departments at the Institute of Philology, namely, the Department of Practical English and the Department of Foreign Languages.

2.2 Instrument

A multiple-choice reading comprehension test was developed by the researchers for this study. The theoretical framework for constructing the test was the subskill-based reading comprehension model advocated by Hughes and Hughes (2020). The subskill-based model of Hughes and Hughes divides reading comprehension skill into two major subskills of expeditious reading and careful reading. Expeditious reading contains fast reading skills such as skimming and scanning and careful reading contains subskills such as understanding main ideas, inference making, and interpreting complex sentences, among others.

Twenty items were written to measure expeditious reading. A long text containing around 3000 words was selected. It contained some sections of the appendices of *The Cambridge Encyclopedia of Language* by David Crystal. Some portions from the index of authors and personalities, the index of topics, the table of the world's languages, the table of special symbols and abbreviations, and the glossary were selected. Questions were constructed based on these selected texts. Sample items include:

- Where is the Chuave language spoken?
- What language family does Nandi belong to?
- On what page can you find information about Black English Vernacular?

The second part of the test was designed to measure careful reading. Items were written to measure inference making, understanding main ideas, and interpreting complex sentences, among others. Twenty items were written to measure careful reading. Ten minutes was allotted for the expeditious section while 20 minutes was given for the careful section.

2.3 Analyses

Four models were estimated. First, a unidimensional model in which all the items measure a single dimension was estimated. A 2-dimensional uncorrelated model in which the 20 expeditious items load on one dimension and the 20 careful items load on a second dimension was estimated. In the next phase, a 2-dimensional correlated model in which the 20 expeditious items load on one dimension and the 20 careful items load on a second dimension was estimated. In this model, the two dimensions were allowed to correlate while in the second model, they were forced to be uncorrelated. In the last phase, a bifactor model was estimated. In this model, a general reading comprehension dimension was assumed (with all the 40 items loading on it) and two uncorrelated dimensions of expeditious and careful reading were also incorporated. In other words, each item simultaneously loads on two dimensions: a general reading comprehension dimension and a specific dimension for the subscale to which the item belongs. The data were analyzed using the ConQuest computer program (Wu et al., 2007).

Table 1 shows the global fit statistics and information criteria (AIC, BIC, CAIC) for the four estimated models. Deviance is -2loglikelihood of the model and the smaller the deviance, the better the model fits (Janssen & De Boeck, 1999). Information criteria are computed based on the

deviance and the number of parameters. The smaller the information criteria, the better the model fits (Akaike, 1974). According to Table 1, the 2-dimensional correlated model has the best fit followed by the bifactor model. The uncorrelated 2-dimensional model has the worst fit as it has the greatest deviance and information criteria values.

Table 1.
Global Fit Statistics and Information Criteria for the Four Models

Model	Deviance	AIC	BIC	CAIC
Uni.	51648	51730	51819	51841
2-dim. Cor.	51137	51237	51211	51239
2-dim. Uncor.	51723	51813	51890	51916
Bifactor	51189	51279	51268	51293

Note: Uni. =unidimensional; 2-dim. Corr.= correlated 2-dimensional; 2-dim. Uncor. =Uncorrelated 2-dimensional

Table 2 shows the reliabilities and variances of the dimensions in the four estimated models. According to Table 2, in the 2-dimensional correlated model, i.e., the best fitting model, the two dimensions of expeditious reading and careful reading have high reliabilities and variances. The reliability in the unidimensional model is the highest. This is obviously because the single dimension of reading comprehension contains 40 items while the separate dimensions of expeditious reading and careful reading each have 20 items. The careful reading dimension has the lowest variance and reliability in the bifactor model.

Table 2.
Variances and Reliabilities for the Dimensions in the Four Models

	Uni.		2-dim. Cor.		2-dim. Uncor.		Bifactor	
Dimension	Var.	Rel.	Var.	Rel.	Var.	Rel.	Var.	Rel.
General Reading	.86	.91	–	–	–	–	.83	.87
Expeditious	–	–	1.08	.85	.83	.79	1.05	.51
Careful	–	–	.83	.81	.72	.74	.08	.11

Note: Uni. =unidimensional; 2-dim. Corr.= correlated 2-dimensional; 2-dim. Uncor. =Uncorrelated 2-dimensional; Var.=variance; Rel.= reliability

3. Discussion and conclusion

In this study, the dimensionality of a reading comprehension test composed of careful reading and expeditious reading items was examined. The reading comprehension model of Hughes and Hughes (2020) was considered. Four models of unidimensional, 2-dimensional with correlated dimensions, 2-dimensional with uncorrelated dimensions and bifactor were estimated.

Global fit statistics including deviance and information criteria showed that the correlated 2-dimensional model has the best fit. Previous research has also shown the multidimensionality of reading comprehension (Zeraatpishe & Sheybani, 2018).

In the bifactor model, where a general factor is estimated along with two specific factors of expeditious reading and careful reading, all the common variance of the items is captured in a general factor and the unique variances are captured in specific dimensions. Table 2 shows that when the common variance is represented in the general factor, a very small amount of variance is left in the careful reading dimension while the variance of the expeditious reading dimension is substantial (Baghaei, 2016). This is an indication that only the expeditious reading ability dimension contains variance above and beyond the general reading comprehension dimensions. The construct of careful reading does not contain much unique variance beyond the general reading comprehension. That is, the general reading comprehension scores represent examinees' careful reading comprehension ability which makes the dimension obsolete. However, after taking into account the common variances, there is still a substantial amount of variance in the expeditious reading dimension. This is an indication that expeditious reading comprehension is a dimension of reading comprehension ability that contains something beyond the reading comprehension ability dimension and needs to be treated separately (Spoden & Fleischer, 2019). That is, examinees' expeditious reading ability should be reported separately from the general reading comprehension ability.

It seems that the ability to read fast and locate information is a distinct construct which deserves its own separate dimension and examinees' scores should be reported separately on this dimension. This could be due to the methods of teaching reading comprehension and the focus that the reading curriculum places on careful reading compared to expeditious reading. Very rarely, at least in the Kazakh education, EFL students are trained in reading long passages to find specific information in a very short time. Most of the emphasis in reading comprehension courses is on careful reading, i.e., analyzing texts and understanding every detail. Perhaps this is the reason why expeditious reading comprehension forms a noticeable second dimension than the careful dimension. However, this does not mean that careful reading is not an important dimension. Since there are only two dimensions here, one could argue that the variance of the careful reading dimension is all absorbed in the general dimension.

Due to the emphasis of the reading curriculums on careful reading, general reading comprehension is almost equated with careful reading. After capturing the variance of the careful reading in the general reading dimension, nothing unique is left in the careful reading dimension. This means that everything in the careful reading dimension is also present in the general reading dimension. But this is not the case for the expeditious reading dimension. After capturing its shared variance with the careful reading dimension, a large portion of unique variance remains in this dimension. This unique variance needs to be taken care of by a separate score to represent examinees expeditious reading ability. This is again another support for the better fit of the correlated 2-dimensional model. That is, examinees' reading ability can best be reported on two separate aspects of careful reading and expeditious reading. With more emphasis and expeditious

reading in future, the distinctness of the two dimensions may disappear in the long run and a single dimension may account for all the variance in reading comprehension items.

The pedagogical implications of the current study are that expeditious reading skills form a separate dimension from careful reading skills (which is basically equivalent to general reading comprehension skill) and needs to be given urgent and full attention in teaching and testing second language reading comprehension. Note that, this does not mean that careful reading and general reading ability are always equivalent. This is a feature of the reading curriculum. When expeditious reading is not given its due attention, it forms a separate dimension which can be distinguished from careful reading. However, in countries where both reading skills are given equal attention in the curriculum, they become one and separation with psychometric models may not be possible.

Future research should focus on the development and measurement of expeditious reading ability. Since expeditious reading takes place with time constraints, IRT models for count data such as the Rasch Poisson Counts Model (RPCM; Rasch, 1960/1980) can be used for modeling the data. For more information on modeling count data with the RPCM see Baghaei and Doebler (2019) and Baghaei et al. (2019). Furthermore, the underlying subskill-based structure of expeditious reading can be evaluated with the linear logistic test model (LLTM) of Fischer (1973). For the application of the LLTM see Baghaei and Ravand (2015), Baghaei and Kubinger (2015), and Hohensinn and Baghaei (2017). Cognitive diagnostic models can also be used to diagnose learners weaknesses and strnegths in expeditious reading which has been neglected by reading teachers and researchers (Alallo et al., 2023; Boori et al., 2023; Effatpanah, et al., 2019; Mohammed et al., 2023).

References

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1–23. doi: 10.1177/0146621697211001

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. doi: 10.1109/TAC.1974.1100705

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. (2016). Modeling multidimensionality in foreign language comprehension tests: An Iranian example. In V. Aryadoust & J. Fox (Eds.), *Trends in Language Assessment Research and Practice: The View from the Middle East and the Pacific Rim* (pp.47-66). Cambridge Scholars.

Baghaei, P. (2012). The application of multidimensional Rasch models in large scale assessment and validation: An empirical example. *Electronic Journal of Research in Educational Psychology, 10*, 233–252.

Baghaei, P., & Aryadoust, V. (2015). Modeling local item dependence due to common test format with a multidimensional Rasch model. *International Journal of Testing, 15*, 71–87. doi: 10.1080/15305058.2014.941108

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. (2015). A cognitive processing model of reading comprehension in English as a foreign language using the linear logistic test model. *Learning and Individual Differences, 43*, 100-105. doi: 10.1016/j.lindif.2015.09.001

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., & 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., 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.

Bond, T., Yan, Z., & Heene, M. (2021). *Applying the Rasch model* (4th Ed.). New York: Routledge.

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

Crystal, D. (1987). *The Cambridge encyclopedia of language*. New York: Cambridge University Press.

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

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 University Press.

Janssen, R., & De Boeck, P. (1999). Confirmatory analyses of componential test structure using multidimensional item response theory. *Multivariate Behavioral Research, 34*, 245–268. doi: 10.1207/S15327906Mb340205

Mohammed, A., Dawood, A. K. S., Alghazali, T., Kadhim, Q., Sabti, A., & Sabit, S. (2023). A cognitive diagnostic assessment study of the reading comprehension section of the Preliminary English Test (PET). *International Journal of Language Testing*, (Special Issue:

Advanced Psychometric Methods in Language Testing), 1-20. doi: 10.22034/IJLT.2022.362849.1195

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Expanded Ed. University of Chicago Press, Chicago, IL. (Originally published 1960, Padagogiske Institut, Copenhagen.)

Reckase, M. D. (2009). *Multidimensional item response theory*. Springer.

Sheybani, E., & Zeraatpishe, M. (2018). On the dimensionality of reading comprehension tests composed of text comprehension items and cloze test items. *International Journal of Language Testing, 8*, 12–26.

Spoden, C., & Fleischer, J. (2019). Multidimensional Rasch models in first language listening tests. In V. Aryadoust & M. Raquel (Eds.), *Quantitative Data Analysis for Language Assessment* (Vol. II) (pp. 33-55). Routledge.

Wu, M. L., Adams, R. J., & Haldane, S. A. (2007). *ACER ConQuest*. Australian Council for Educational Research.