Language Testing in Asia

RESEARCH Open Access

Rasch testlet model and bifactor analysis: how do they assess the dimensionality of large-scale Iranian EFL reading comprehension tests?

Masoud Geramipour

Correspondence: mgramipour@khu.ac.ir
Department of Curriculum Studies and Educational Research, Kharazmi University, No. 43, South Mofatteh Ave, Tehran, Iran

Abstract

Rasch testlet and bifactor models are two measurement models that could deal with local item dependency (LID) in assessing the dimensionality of reading comprehension testlets. This study aimed to apply the measurement models to real item response data of the Iranian EFL reading comprehension tests and compare the validity of the bifactor models and corresponding item parameters with unidimensional and multidimensional Rasch models. The data collected from the EFL reading comprehension section of the Iranian national university entrance examinations from 2016 to 2018. Various advanced packages of the *R system* were employed to fit the Rasch unidimensional, multidimensional, and testlet models and the exploratory and confirmatory bifactor models. Then, item parameters estimated and testlet effects identified; moreover, goodness of fit indices and the item parameter correlations for the different models were calculated. Results showed that the testlet effects were all small but non-negligible for all of the EFL reading testlets. Moreover, bifactor models were superior in terms of goodness of fit, whereas exploratory bifactor model better explained the factor structure of the EFL reading comprehension tests. However, item difficulty parameters in the Rasch models were more consistent than the bifactor models. This study had substantial implications for methods of dealing with LID and dimensionality in assessing reading comprehension with reference to the EFL testing.

Keywords: EFL testing, Local item dependence, Test dimensionality, Rasch testlet model, Bifactor analysis, Unidimensional and multidimensional Rasch models

Introduction

Background of study

Linguistic ability is a very complicated and interrelated attribute, which simultaneously needs different proficiencies (Wainer and Wang, 2000). Students' reading comprehension achievement is dependent upon the accomplishment of several cognitive skills at the word, sentence, and the whole test level. Therefore, recognizing the construct of the reading comprehension test is helpful to identify the progressions in English as a foreign language (EFL) testing (Schindler et al., 2018). Construct validity is at the heart

of the test validity notion (Messick, 1989). One of the important features of construct validity is the test dimensionality, which is very important in the development and validation of tests of second language (L2) ability (Dunn & McCray, 2020). Test dimensionality, which is defined as the minimum number of examinee abilities measured by the test items, is a unifying concept that underlies some of the most essential issues in the development of large-scale tests (Tate, 2002). Advanced measurement techniques are capable to quantify the test construct validity and dimensionality of large-scale assessments (Reder, 1998; Westen & Rosenthal, 2003). In doing so, the Rasch testlet (Wang & Wilson, 2005) and bifactor analysis models (Holzinger and Swineford, 1937; Reise, 2012; Schmid & Leiman, 1957) are some new advanced methods, which are capable of investigating the construct dimensionality of the EFL reading comprehension tests.

Reading comprehension assessments usually consist of various text passages, each following some items which are grouped into item bundles (Rosenbaum, 1988). In this context, the item bundle that shares a common reading comprehension passage is a testlet. Moreover, testlets may refer to any common stimuli such as a graph, table, diagram, map, item stems, and scenario (Wang et al., 2005).

Local item dependency (LID) is the main property of testlets, which befalls when a special dependency exists between items (Wilson, 1988). The testlet-based process is exercised in EFL reading comprehension testing. For example, some EFL large-scale tests such as the Test of English as Foreign Language (TOEFL) from Educational Testing Service (ETS), the Cambridge exams of the University of Cambridge Local Examinations Syndicate (UCLES), the First Certificate of English (FCE), and International English Language Testing System (IELTS) all offer passages to examinees to read and answer the following related questions (Chalhoub-Deville & Turner, 2000).

Standard item response theory (IRT) models may not function properly with English language reading testlets and lead to small distortion of parameter estimates and overestimated reliability (Li et al., 2010). This is because of overlooking of the testlet effects, which is different from test to test (DeMars, 2012). Testlet effect is a random effect variance induced by the LID of the test. The more the variance, the more the effect contained inside the testlet (Wainer and Wang, 2000). Testlet response theory (TRT) (Wainer et al., 2007) and the bifactor models are newly developed to model the testlet effects in this regard. Moreover, the multidimensional item response theory (MIRT) models can be also used to address the testlet effect (Baghaei, 2012; Wainer et al., 2007).

Rasch models provide the possibility of fundamental measurement in standard testing (Andrich, 1988). The models have extensive applications in objective measurement of dichotomous or polytomous variables in human sciences, by applying either unidimensional or multidimensional Rasch models (Bond & Fox, 2015). Moreover, the theory of Rasch model has been gradually increased in language assessment through past decades (Fan et al., 2019; Aryadoust et al., 2020), while the Rasch model of the TRT, as a special case of the bifactor analysis model, was introduced and developed by Wang and Wilson (2005).

Testlets have extensive use in EFL testing, and as the face and the nature of the testlets are different from the other forms of testing, they need relevant methods of analyzing test validity and dimensionality. Moreover, an investigation of the testlet effect in

EFL reading comprehension tests is paramount accordingly. Thus, the present study aimed to apply and compare the Rasch testlet and bifactor models on the Iranian large-scale EFL reading comprehension tests to investigate the testlet effects and to compare its validity to the other relevant versions of unidimensional and multidimensional Rasch models.

Rasch unidimensional, multidimensional and testlet models, and the bifactor analysis

Nowadays, IRT models have been widely used to test scoring in large-scale educational testing. Rasch model is a variation of the IRT models. Especially in the field of language testing, Rasch measurement has been frequently used in the assessment of reading, writing, speaking, and listening skills (Aryadoust et al., 2020). It is an item analysis model with logistic item characteristic curves of equal slope (Andersen, 1973). Reckase (2009) extended the standard Rasch model to the multidimensional Rasch model, in which, the probability of answering an item correctly is the function of more than one latent ability at the same time.

Local independence of test items is the common assumption in all IRT models. It means that the examinee abilities provide all information needed to explain performances, and controlling for the trait level, all of the other affecting factors are random. This assumption is violated when IRT models applied on tests including testlets, because answering to pairs of items are correlated given the ability (Wainer et al., 2000).

TRT is a solution to model the dependency of the test items, when we confront the problem of local dependency (Baldwin, 2007). Extending the Birnbaum (1968) three parameters logistic (3PL) IRT model, Wainer et al. (2000) added the testlet effect to the model and developed the 3PL TRT model. A path diagram for an assumptive testlet model, just like the model that is presented in this study, is shown in Fig. 1.

In the Rasch testlet model, discrimination parameters (loading) are constrained to be equal in each testlet (Wang & Wilson, 2005) (see the Appendix for the technical formulas of the unidimensional, multidimensional, and the Rasch testlet models).

Another solution to model the dependency among test items in testlets such as a reading test is a bifactor model (Rijmen, 2010). The bifactor model, also known as a hierarchical model (Markon, 2019) or even a nested-factor model (Brunner et al., 2012), is a latent structure in which the items load on a general factor on the one hand and show the factor structure of specific factors for each item on the other. The general factor accounts for the total variation among items and shows what items have in common, and specific factor structure is interpreted like the other ordinary factor analytic methods (Reise et al., 2010).

Moreover, two major variations of exploratory and confirmatory analysis are available for structuring the specific factors in the bifactor model (Reise, 2012). A template for exploratory and confirmatory bifactor model with an assumptive general factor (e.g., reading ability) and three specific factors (e.g., decoding, fluency, and vocabulary skills) in which three items are nested is illustrated in Fig. 2.

As shown in Fig. 2, in the exploratory bifactor model, item factor loadings on all of the specific factors are freed to be estimated, whereas they are only constrained to be related to the specific factors in the confirmatory bifactor model. It is worth noting that the Rasch testlet model is actually a confirmatory bifactor analysis model in which the

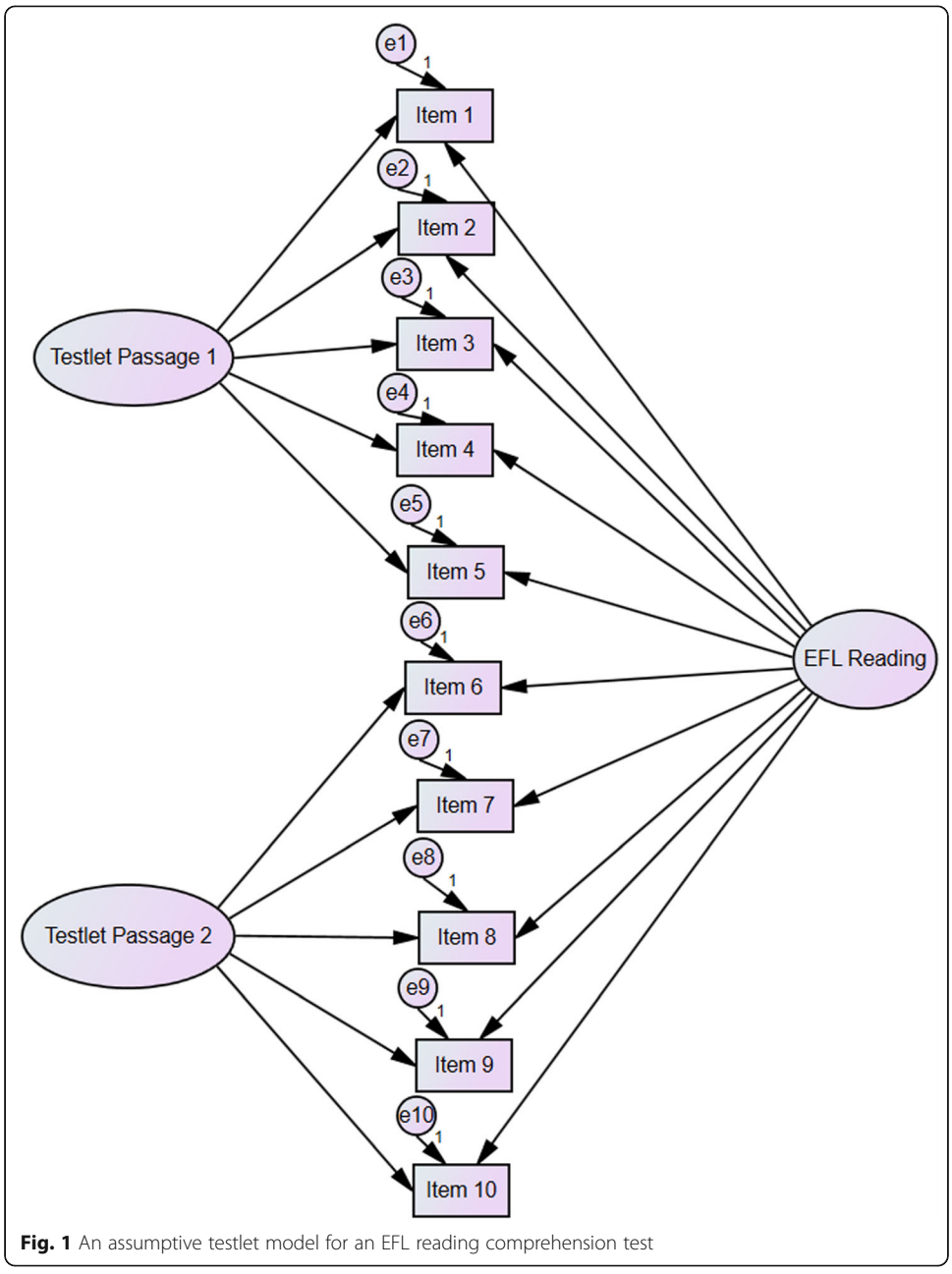


Fig. 1 An assumptive testlet model for an EFL reading comprehension test

factor loadings on the specific factors are restricted to be proportional to the loadings on the general factor within each testlet (Li et al., 2006). That is, the Rasch testlet model is actually the Rasch version of a bifactor model (Jiao et al., 2013). Therefore, in addition to the part b of the Fig. 2, the general structure of Fig. 1 is also true for the confirmatory bifactor analysis model, where the magnitudes of factor loadings are just estimated in each specific factor.

Literature review

The test dimensionality refers to the number of latent variables that are measured by a set of test items. Therefore, an essential unidimensional test measures predominantly one latent variable, while a multidimensional test measures more than one latent

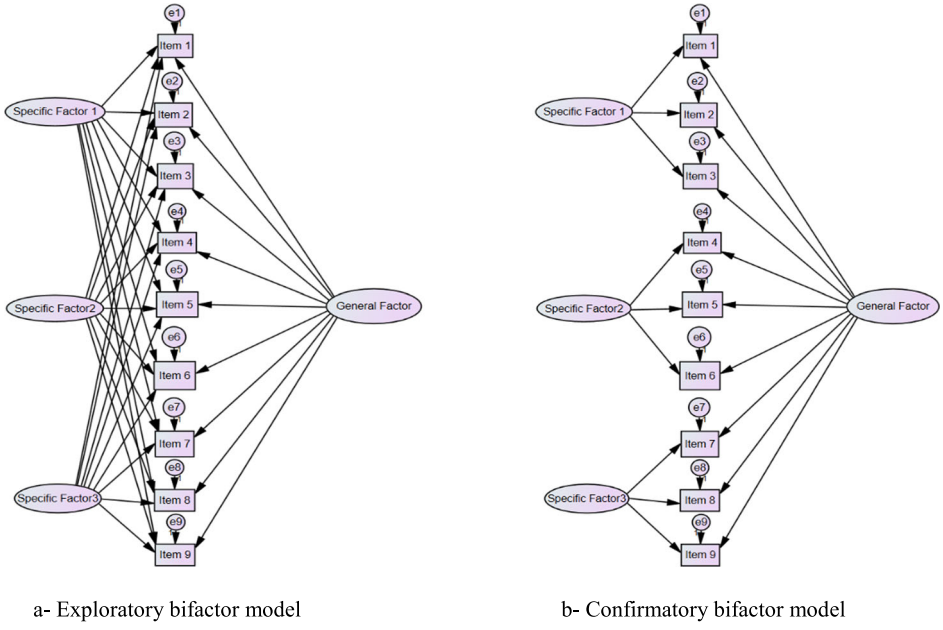


Fig. 2 Illustrative bifactor models with three specific factors and three items within each factor

variable (Mellenbergh, 2019). Dimensionality in language tests is directly related to the construct validity (Henning, 1992). To reach the goal, an evidence-based process of determining the total scale of a test and its subscales is essential to provide an argument for construct validity. This is exactly what dimensionality analysis does to provide construct validity argument for a test (Slocum-Gori & Zumbo, 2011).

The unidimensional Rasch models have been applied to assess the dimensionality of EFL reading tests for decades (Aryadoust et al., 2020). In part of a more general study on applying unidimensional IRT models on EFL vocabulary and reading tests, Choi and Bachman (1992) applied the one-parameter Rasch model on the reading comprehension section of the TOEFL and FCE. Using Stout approach for assessing unidimensionality (Stout, 1987), they showed that the reading tests are not unidimensional at 5% confidence interval. They also determined that the unidimensional Rasch model tends to be significantly less fitted to the item response data in comparison to 2PL and 3PL IRT models. In another research of that type, Boldt (1992) modified the Classical Rasch model through estimating and fixing the guessing parameter to a certain magnitude and applied it to the TOEFL listening, structure, and reading comprehension sections. It was initially expected that the 3PL IRT model was more efficient, but they reported that the Rasch model could equally predict the person's success on the items and the model is competitive to the 3PL IRT model. Moreover, Lee (2004) investigated local item dependency (LID) of a Korean EFL reading comprehension test by IRT-based Q_3 index (Yen, 1984). He demonstrated clear evidences of passage-related LID of the 40 items test. It was found that positive LID existed among items within testlets which induced the passage content. Not in the EFL testing occasion, Monseur et al. (2011) evaluated the LID of the reading component data of Programme for International

Student Achievement (PISA[1]) (2000 and 2003) through the IRT-based Yen's Q_3 method. They reported a moderate testlet effect; however, the global context dependencies were clear for a large number of reading comprehension sections in different countries. As the application of unidimensional IRT models on the EFL reading tests reviewed above, not all of the aforementioned studies went beyond assessing the unidimensionality assumption, where in the meantime, multidimensional IRT methods were growing in the field of EFL language testing (e.g., Ackerman, 1992; McKinley & Way, 1992).

In more advanced unidimensional Rasch analysis studies, Baghaei and Carstensen (2013) fitted mixed Rasch model to an EFL high school reading comprehension test consisting of short and long passages with total number of 20 items. They reported that the model fitted significantly better to the item response data than the standard Rasch model. The model also indicated that the students were divided into two classes of high proficient students in short and long passages. Furthermore, Aryadoust and Zhang (2016) fitted the mixed Rasch model to a large sample of Chinese college students taking an EFL reading comprehension test. They found that 48 out of 50 items of the EFL reading test are well fitted to the mixed Rasch model. Moreover, the results indicated two distinct latent classes within students, in which one class is good at reading in depth and the other performs better in skimming and scanning text passages. The act of analyzing a measure requires a number of essential assumptions. The most important among these assumptions is that the construct is unidimensional (Briggs & Wilson, 2003). Although Baghaei and Carstensen (2013) and Aryadoust and Zhang (2016) applied advanced measurement analysis on the EFL reading tests, they could not be able to reject the multidimensionality assumption of the EFL reading test in any way.

The application of multidimensional Rasch model on EFL reading tests has also some case studies, as follows: Baghaei (2012) applied a compensatory multidimensional item response theory (MIRT) model on an English comprehension test including two listening and reading comprehension tests. Each test had also two subtests, which measured informational and interactional listening skills and expeditious and careful reading skills, respectively. Goodness of fit of a unidimensional and two multidimensional Rasch models were evaluated by Akaike's information criterion (AIC) (Akaike, 1974) and Bayesian information criterion (BIC) (Schwarz, 1978). The results generally supported the multidimensional Rasch models with two and four dimensions against unidimensional Rasch model, whereas the four-dimensional model encompassing the aforementioned subtests was superior to two-dimensional model including only the listening and reading dimensions. Moreover, Baghaei and Grotjahn (2014) analyzed an English C-test[2] including two spoken and written discourse passages through confirmatory multidimensional Rasch analysis. They compared the results with the standard unidimensional Rasch model, and they illustrated that the two-dimensional model is the best fitting confirmatory model for the C-test. Using multidimensional IRT models improve the precision of measurement in tests substantially, especially when the test is

[1]It is a project that is held regularly every 3 years by the OECD (Organization for Economic Cooperation and Development)

[2]A kind of cloze test consisting of several short passages in which the second half of every second word is deleted

short and includes more than one unidimensional test (Wang et al., 2004). However, the performance of multidimensional IRT models to deal with test multidimensionality and testlet effects in comparison to other counterpart models such as the bifactor models has not well researched yet.

In comparison to the unidimensional and multidimensional models, employing TRT models to analyze EFL reading tests are more recent. As pioneers of the Rasch testlet model, Wang and Wilson (2005) applied the Rasch testlet model on a Taiwanese high school English test with 44 items and 11 testlets. The likelihood deviance (− 2 log-likelihood) of the model was significantly smaller than unidimensional Rasch model, indicating the local dependence existed among items within the testlets. Of course, some fluctuations between empirical and expected response curves were observed because of the large sample size of 5000 students; however, the Rasch testlet model was fairly well fitted to the item response data. In another research, Baghaei and Ravand (2016) studied the magnitudes of local dependencies generated by EFL cloze and reading test passages using the 2PL TRT model. They showed that the 2PL TRT model is better fitted to the EFL reading testlet data than the counterpart 2PL IRT model; however, the testlet effect for the reading test was ignorable. Rijmen (2010) also applied three multidimensional models on a testlet-based data stemming from an international English test. He showed that the 2PL TRT model performed better than a unidimensional 2PL IRT model; however, the deviance of the bifactor model was better than the TRT/second-order model. Moreover, in an advanced two-level study, Ravand (2015) applied the TRT model to assess the testlet effect in a reading comprehension test held for Iranian applicants for English master's program in state universities. Four testlets were investigated, where half of the testlet effects were negligible according to criteria proposed by either empirical or simulation studies (Glas et al., 2000; Wang et al., 2002; Wang & Wilson, 2005; Zhang et al., 2010; Zhang, 2010, b). It was also found that ignoring local dependence would result in overestimation in lower and upper bounds of the ability continuum, even if the item difficulty parameters were the same as the conventional models. TRT models were typically fitted against unidimensional IRT models in studies such as Baghaei and Ravand (2016), Ravand (2015) and Wang and Wilson (2005) studies. Although testlet effects were evaluated in the aforementioned studies; however, the goodness of fit of the TRT models were not examined to the multidimensional or bifactor models alongside with the conventional IRT models.

Not in the EFL testing but in a similar testing occasion, Eckes and Baghaei (2015) addressed the issue of local dependency of C-tests in a German as foreign language test using fully Bayesian 2PL TRT model. As they highlighted, when local dependency of the C-test is ignored, reliability of the test is slightly overestimated, whereas the testlet effects of eight texts of the C-test were entirely ignorable. Nevertheless, in contrast to the previous studies, the results showed that the precision of the 2PL TRT model is less than the conventional 2PL IRT model. In another study in the context, Huang and Wang (2013) applied hierarchical testlet model and hierarchical IRT model to some ability and non-ability tests including a Taiwanese EFL reading test and compared the models to the unidimensional 1PL, 2PL, and 3PL IRT models. BIC indices showed that the testlet model fitted significantly better than other aforementioned models. They also showed that ignoring testlet effect led to biased estimation of item parameters, underestimation of factor loadings, and overestimation of the test reliability. Kim

(2017) employed the TRT model to investigate testlet effect in an eighth-grade reading comprehension test in the USA. Using a χ^2 for evaluating local independence, he revealed that there were significant dependencies among the test items, where the testlet effects were evaluated high with respect to the magnitudes of the testlet item dependencies. He also compared the goodness of fit of the TRT model with the unidimensional IRT model by AIC and BIC indices in his Ph.D. dissertation, where he concluded that the TRT model fitted better to the test. Jiao et al. (2012) applied an advanced multilevel testlet model for dual local dependence on a state reading comprehension test for high school graduation in the USA. Using The Deviance Information Criterion (DIC), they found that the proposed model and the multilevel model are better fitted than the testlet model to four testlet response data, whereas the testlet model showed better DIC in comparison to standard Rasch model. Moreover, the results of the advanced and typical testlet models showed that the estimated testlet variances for all reading testlets were small. Chang and Wang (2010) fitted standard IRT and TRT models to 10 reading testlets of the Progress in International Reading Literacy Study (PIRLS) test held in 2006. As expected, TRT fitted significantly better to PIRLS (2006) data in comparison to the unidimensional IRT model. They also reported a testlet variance ranged from .168 to .489, which indicated negligible to moderate testlet effect in the test. Moreover, Jiao et al. (2013) applied the Rasch testlet model and a three-level one-parameter testlet model to a large-scale assessment k-12[3] reading test battery including six testlet passages from grades 9 to 11. Applying different methods of Bayesian and non-Bayesian estimation, they reported that half of the testlet effects were negligible for the reading test passages. looking at the findings of the above research, it becomes apparent that not only the testlet model perform better in EFL reading comprehension testing, but also it has the same function in other non-EFL situations of testing reading. Moreover, the testlet effects in testing non-EFL reading comprehension look as random as testing reading in EFL examinations. However, in none of the reviewed studies, the bifactor models were employed or compared to the TRT models.

To review some studies on the application and comparison of bifactor model to investigate the dimensionality of EFL reading comprehension tests, Wu and Wei (2014) compared the bifactor model to the IRT and TRT models to investigate the testlet effect in an EFL passage-based test in China. The results showed that there was not a high degree of dependency between the passage items. They also observed that the item difficulty parameters were the same, while the item discriminations were very different. There were also similarities between ability estimations in the bifactor model with the other models; however, a considerable discrepancy was observed in the standard errors of the different models. Byun and Lee (2016) investigated testlet effect in a Korean EFL reading comprehension test using the MIRT bifactor model. They found that the bifactor model is the best fitting model, and passage topic familiarity is not necessarily related to the factor scores. They also reported that the overall topic familiarity is correlated to the general reading ability score of the examinees. Moreover, in a recent study in the field of language testing, Dunn and McCray (2020) applied and compared

[3]An American expression that indicates the years from kindergarten to 12th school grade, including the years supporting primary and secondary education in the United States before college. Several other countries follow the same K-12 educational system such as Afghanistan, Australia, Canada, Ecuador, China, Egypt, India, Iran, the Philippines, South Korea, and Turkey

a range of CFA model structures on data sets from British Council's Aptis[4] English test (O'Sullivan, and Council, B., 2012). Using some absolute and relative measures of goodness of fit (Sun, 2005), they successfully showed that how the bifactor model spearheads other confirmatory factor analysis measurement models in the field of second language testing. However, in all of the aforementioned studies, just linear models of confirmatory factor analysis were considered to be applied on language test data and compared with the bifactor model.

In a different testing situation, to explore the utility of the bifactor model, Betts et al. (2011) investigated the early literacy and numeracy measured by the Minneapolis Kindergarten Assessment (MKA) tool in the USA to correct for the anomalies found in the factor structure of MKV in the previous research literature. Results of the study showed that the bifactor model provides a strong model conceptualization tool and a precise predictive model for later reading and mathematics. In another study in a non-EFL testing occasion, Foorman et al. (2015) explored the general and specific factors of an oral language and reading test using the bifactor analysis model. Results supported a bifactor model of lexical knowledge of reading rather than a simple view of a three-factor model including decoding fluency, vocabulary, and syntax. Finally, in a much more similar research to this study, Min and He (2014) examined the relative effectiveness of the multidimensional bifactor and testlet response theory models in dealing with local dependency in reading comprehension testlets of the Graduate School Entrance English Exam (GSEEE) in China. They concluded that although the bifactor model is better fitted than the TRT model to the data, but both models produce the same results in terms of item and ability parameters. Moreover, they found that the unidimensional IRT models did not fit to the item response data and had a bigger impact on item slopes other than the item and person parameters. However, the MIRT and different approaches of exploratory and confirmatory bifactor models were not considered for further explanation of the test dimensionality and testlet effects.

As reviewed above in the research literature, the results about the magnitudes of testlet effect or even the existence of such effects in EFL or non-EFL reading comprehension passages were contradictory. Some studies concluded that the testlet effects were significant (e.g., Kim, 2017) or moderate (e.g., Monseur et al., 2011), whereas the other findings indicated that the testlet effects for the reading passages were small (e.g., Jiao et al., 2012, Wu & Wei, 2014), negligible, or even some testlets of the whole reading comprehension section lacked the effect (e.g., Baghaei & Ravand, 2016; Chang & Wang, 2010; Jiao et al., 2013; Ravand, 2015). Therefore, it seems that more research needed to conclude about the random nature of the testlet effects in reading passages, especially in the EFL testing context, to judge that the effects are indigenous in the reading testlets, or it is a random effect that is rooted to the occasion or content of testing.

The other important features of the Rasch testlet model and the bifactor analysis are the goodness of fit and item and ability parameters produced by the models in comparison to the other conventional and multidimensional Rasch models. Typically, TRT models were at least compared to either the unidimensional Rasch model or 2PL/3PL

[4]The Aptis test is a flexible English language test that helps an organization or an institution to measure all four English skills including reading, writing, listening, and speaking together with the core mandatory component. It is also possible that only one skill, e.g., listening, being assessed in the Aptis test (O'Sullivan, and Council, B., 2012)

IRT models in most of the past research (e.g., Baghaei, 2015; Byun & Lee, 2016; Chang & Wang, 2010; Huang & Wang, 2013; Jiao et al., 2012; Kim, 2017; Min & He, 2014; Ravand, 2015; Rijmen, 2010; Wang & Wilson, 2005; Wu & Wei, 2014). All of the aforementioned studies except one (Baghaei and Aryadoust, 2015) favor TRT models in terms of goodness of fit; however, the results of item and ability parameters calibration were not highly consistent in the applied TRT models. Therefore, the correlation of item parameters in the Rasch models need to be more scrutinized, especially with real data sets of large-scale assessments.

Moreover, multidimensional IRT models were also investigated against conventional IRT models in some case studies (e.g., Baghaei, 2012; Baghaei & Grotjahn, 2014; Byun & Lee, 2016). In Rasch models family, Wang and Wilson (2005) only compared Rasch testlet model to the unidimensional Rasch model. However, due to the popularity of the Rasch models among language testing specialists, the validity of the Rasch testlet model has not yet been compared with the multidimensional Rasch model or the bifactor analysis model with real data sets of EFL reading comprehension tests. Therefore, following some recent research on the EFL reading comprehension testing in the Iranian large-scale university entrance examinations (e.g., Geramipour & Shahmirzadi, 2018; Geramipour & Shahmirzadi, 2019; Geramipour, 2020), present research also aims to introduce a new collection of dimensionality analysis methods to the field of language testing.

Research questions

Thus, as the main purpose of the current study is to apply the Rasch testlet and bifactor models to analyze the LID and dimensionality of the reading comprehension tests and compare them to the unidimensional and multidimensional Rasch models, the questions that guided this study are the following:

1. How large are the testlet effects for the EFL reading comprehension tests?
2. Is there any correlation between item parameters (difficulty indices) of the Rasch testlet and bifactor models with each other and the other unidimensional and multidimensional Rasch models in the EFL reading comprehension tests?
3. Are the Rasch testlet and bifactor models better fitted to the item response data of the EFL reading comprehension tests in comparison to the other unidimensional and multidimensional Rasch models?

Method

Instruments

The EFL reading comprehension section of the Iranian national university entrance examination was analyzed in this study from 2016 to 2018. The test is composed of 2 reading passages with 10 multiple choice items which are part of a high-stakes test held annually to admit the candidates to Ph.D. programs in English Language studies. The test is designed for students with a master's degree who aim to pursue education in Ph.D. degree in state universities. The test measures the knowledge of general English consisting of three sections of grammar (8 items), vocabulary (12 items), and the reading comprehension section (10 items) which was chosen for the Rasch and bifactor analysis. The reading comprehension section includes 2 passages with 10 items evenly distributed in each passage.

Participants

The population data of the reading section were provided by the Iranian national organization of educational testing in the format of Excel data files. Then, sample sizes of 4200 (61.30% men and 38.70% women), 3220 (60.40% men and 39.60% women), and 4500 (67.20% men and 32.80% women) were randomly selected from the Excel data files of 2016, 2017, and 2018 examinations respectively. In doing so, one booklet of the two presented booklets to the examinees was randomly selected each year; then, random cluster samples were proportionately drawn from the population based on the participants' gender. IRT analysis methods require item response data from large sample sizes around 1000 examinees or more to yield accurate item and ability parameters (Hambleton, 1989). In addition, there is no shortage of recommendations regarding to the sample size needed to do factor analysis. Absolute ranges of 100 over 1000 examinees are often suggested for conducting factor analysis (Mundfrom et al., 2005). Therefore, the sample size in the present study is a strength.

Data analysis

At last, different packages of *R system* (R Core Team, 2019) were employed for the sake of statistical data analysis. Unidimensional Rasch analysis was done by *ltm* package (Rizopoulos, 2006), multidimensional Rasch models was fitted to the data through *mirt* package (Chalmers, 2012), and the Rasch testlet and the bifactor models were simultaneously run by *TAM* package (Robitzsch et al., 2019) and *sirt* package (Robitzsch, 2019). Moreover, unidimensionality of the EFL reading comprehension tests were evaluated by the unified parallel analysis method (Drasgow & Lissak, 1983) and the *ltm* package in advance of the main analysis.

Results

Before applying the Rasch models and answering the research questions, unidimensionality of the test data under the unidimensional Rasch models was inferentially analyzed. The results of checking unidimensionality of EFL reading comprehension sections using the *ltm* package of the *R* software are shown in Table 1. Moreover, Fig. 3 shows the Eigenvalue plots derived from the observed and simulated item response data for testing unidimensionality through the unified parallel analysis.

As shown in Table 1 and Fig. 3, the significant differences between the observed and Monte Carlo simulated Eigenvalues show that the unidimensionality assumption holds for none of the EFL reading comprehension tests. Thus, there is a strong reason to use multidimensional measurement models including the Rasch testlet and bifactor models to explain the dimensionality of the EFL reading comprehension tests.

Table 1 Testing unidimensionality of the EFL reading comprehension tests using the unified parallel analysis method

Statistic	Examination year		
	2016	2017	2018
Second Eigenvalue in the observed data	0.89	0.64	0.47
Average of second Eigenvalues in Monte Carlo samples	0.23	0.21	0.17
Monte Carlo samples	100	100	100
p value	0.001	0.001	0.001
Result of testing unidimensionality	Rejected	Rejected	Rejected

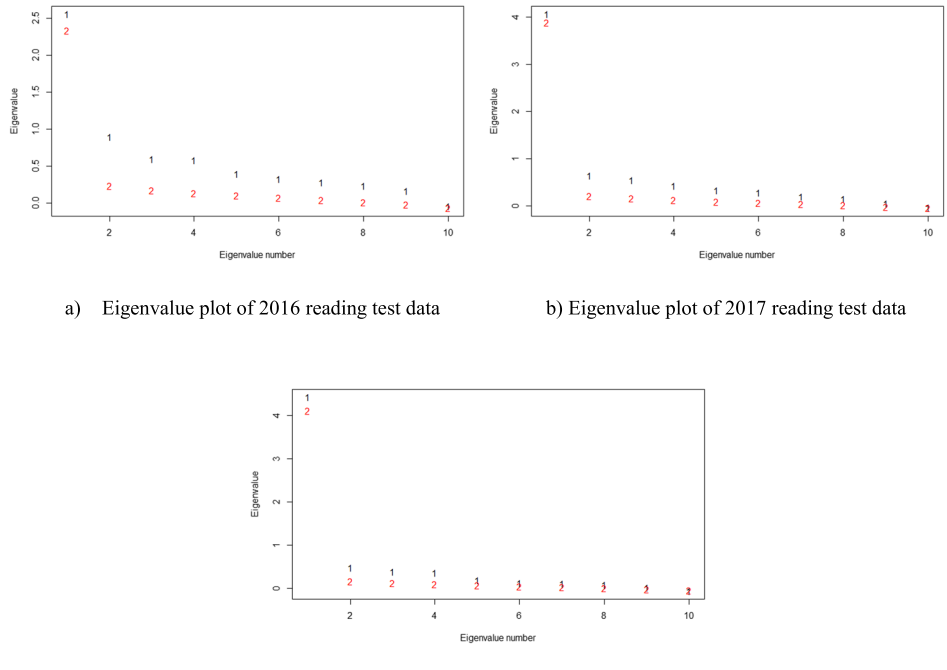


Fig. 3 Eigenvalue plots produced by the unified parallel analysis for the reading comprehension data

How large are the testlet effects for the EFL reading comprehension tests?

Table 2 shows the testlet effect variances estimated by the Rasch testlet model using the *TAM* and *sirt* packages of the *R* software for each EFL reading comprehension testlets from the Iranian 2016 to 2018 examinations.

As shown in Table 2, no substantial testlet effect is observed among testlet passages based on the criteria (values close to 1) proposed by Glas et al. (2000). However, except for one testlet, all of the other testlet effects were higher than 0.25, and then non-negligible according to criteria proposed by Glas et al. (2000), Wang and Wilson (2005), and Zhang et al. (2010).

Is there any correlation between item parameters (difficulty indices) of the Rasch testlet and bifactor models with each other and the other unidimensional and multidimensional Rasch models in the EFL reading comprehension tests?

Item calibration (difficulty and discrimination parameters) results of applying Rasch unidimensional, multidimensional, and testlet models and the exploratory and confirmatory bifactor models using all of the aforementioned *R* packages are briefed in Table 3.

Table 2 Testlet effect variances per EFL reading testlet from the 2016 to 2018 examinations

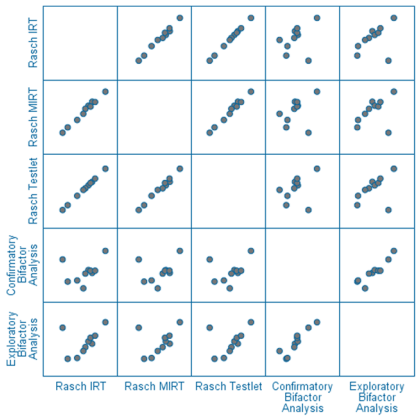
Passage	Examination		
	2016	2017	2018
Testlet 1	0.63	0.25	0.22
Testlet 2	0.32	0.39	0.40

Table 3 Item difficulty and discrimination parameters under different Rasch and bifactor models

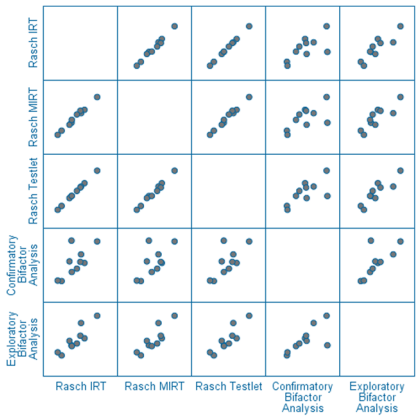
Items/testlets		Models/parameters													
		Rasch IRT		Rasch MIRT		Rasch testlet		Confirmatory bifactor				Exploratory bifactor			
		a	*b*	*a*	*b*	*a*	*b*	*b*	GF[a]	F1	F2	*b*	GF	F1	F2
2016 Reading Comprehension Examination															
1	Testlet 1	0.99	1.51	0.61	1.60	0.91	1.45	1.72	0.49	0.43	-	1.74	0.46	0.06	0.47
2		0.99	2.45	0.61	2.58	0.91	2.35	2.39	0.47	0.03	-	2.32	0.33	0.20	0.17
3		0.99	2.30	0.61	2.43	0.91	2.21	2.61	0.65	0.06	-	2.41	0.43	0.34	0.14
4		0.99	2.19	0.61	2.31	0.91	2.10	2.34	0.55	0.20	-	2.19	0.34	0.38	0.00
5		0.99	1.31	0.61	1.39	0.91	1.27	3.43	0.59	0.73	-	2.94	0.60	− 0.01	0.70
6	Testlet 2	0.99	2.39	0.52	2.40	0.95	2.31	2.56	0.54	-	0.25	2.55	0.45	0.35	0.15
7		0.99	2.56	0.52	2.57	0.95	2.48	2.56	0.40	-	0.32	2.63	0.35	0.42	− 0.02
8		0.99	2.11	0.52	2.12	0.95	2.04	1.20	0.34	-	0.25	2.04	0.29	0.36	− 0.03
9		0.99	2.96	0.52	2.97	0.95	2.86	4.09	0.30	-	0.72	3.15	0.36	0.47	− 0.07
10		0.99	1.87	0.52	1.88	0.95	1.81	1.81	0.30	-	0.35	1.77	0.28	0.32	− 0.01
2017 Reading Comprehension Examination															
1	Testlet 1	1.52	1.15	0.70	1.66	0.96	1.71	2.94	0.72	0.54	-	1.76	0.62	0.09	0.25
2		1.52	1.65	0.70	2.41	0.96	2.45	2.08	0.40	0.22	-	2.03	0.38	0.04	0.17
3		1.52	2.14	0.70	3.15	0.96	3.17	2.91	0.53	− 0.15	-	2.90	0.48	0.01	0.26
4		1.52	1.07	0.70	1.55	0.96	1.59	2.14	0.81	− 0.10	-	1.93	0.72	0.12	0.27
5		1.52	1.52	0.70	2.21	0.96	2.25	2.41	0.70	− 0.06	-	2.62	0.64	− 0.04	0.39
6	Testlet 2	1.52	1.49	0.84	2.35	0.94	2.20	2.12	0.58	-	0.19	2.13	0.57	0.16	0.15
7		1.52	0.77	0.84	1.20	0.94	1.13	1.38	0.68	-	0.37	1.37	0.70	0.31	0.07
8		1.52	0.62	0.84	0.96	0.94	0.91	1.40	0.71	-	0.49	1.49	0.76	0.44	− 0.03
9		1.52	1.16	0.84	1.83	0.94	1.71	1.73	0.57	-	0.33	1.73	0.59	0.27	0.05
10		1.52	1.36	0.84	2.14	0.94	2.00	1.86	0.47	-	0.33	1.83	0.50	0.24	0.04
2018 Reading Comprehension Examination															
1	Testlet 1	1.53	0.47	0.75	0.70	0.96	0.71	0.64	0.46	0.32	-	0.63	0.50	− 0.04	0.23
2		1.53	0.79	0.75	1.18	0.96	1.18	1.42	0.72	0.29	-	1.42	0.73	0.04	0.24
3		1.53	1.04	0.75	1.56	0.96	1.57	1.39	0.49	0.24	-	1.40	0.51	0.00	0.20
4		1.53	1.11	0.75	1.66	0.96	1.67	1.89	0.69	0.27	-	1.88	0.71	0.06	0.21
5		1.53	0.74	0.75	1.10	0.96	1.11	1.15	0.61	0.31	-	1.16	0.64	− 0.01	0.26
6	Testlet 2	1.53	1.16	0.88	1.89	0.94	1.74	1.86	0.69	-	0.14	1.87	0.66	0.25	0.01
7		1.53	1.19	0.88	1.94	0.94	1.78	1.95	0.72	-	0.08	1.97	0.68	0.25	0.01
8		1.53	0.36	0.88	0.58	0.94	0.53	0.63	0.75	-	− 0.14	0.58	0.68	0.17	0.09
9		1.53	0.89	0.88	1.45	0.94	1.33	2.22	0.85	-	0.26	2.07	0.81	0.32	0.00
10		1.53	2.19	0.88	3.56	0.94	3.26	2.83	0.46	-	0.08	2.83	0.44	0.15	0.02

[a]*GF* General factor

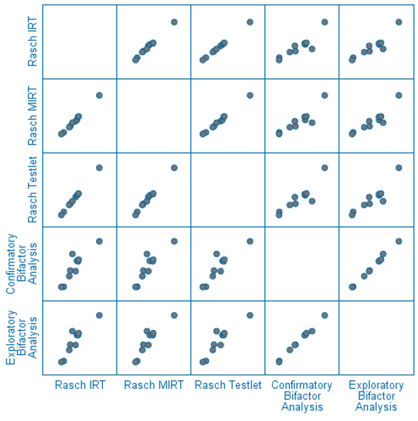
As shown in Table 3, seemingly, item difficulty parameters show a very high positive (almost perfect) level of consistency within the different Rasch models in the 3 years of examinations. Separate matrix scatter plots (a, b, and c) of the item difficulty parameters (b) under unidimensional Rasch model (IRT), multidimensional Rasch model (MIRT), the Rasch testlet model, and confirmatory and exploratory bifactor models for 3 years of the examinations are depicted in Fig. 4.



a) The 2016 reading comprehension exam.



b) The 2017 reading comprehension exam.



c) The 2018 reading comprehension exam.

Fig. 4 Matrix scatter plot of item difficulty parameters in the applied Rasch and the bifactor models from the 2016 to 2018 examinations

As seen in parts a, b, and c of the Fig. 4, item difficulty parameters in the applied Rasch models are very highly correlated, whereas the magnitudes of the correlation coefficients range from ρ = 0.99 to ρ = 1. It is worth noting that the item difficulties of the Rasch testlet model are somehow more positively correlated with the unidimensional Rasch model (ρ = 1) than the multidimensional IRT model (ρ = 0.99). However, the item difficulty correlations for the bifactor models were not as perfect as the Rasch models.

Is the Rasch testlet model better fitted to the item response data of the EFL reading comprehension tests in comparison to the other unidimensional and multidimensional Rasch models and the bifactor analysis model?

Finally, to answer the last research question, the goodness of fit indices of the applied Rasch models to the EFL reading comprehension tests were systematically evaluated by log-likelihood statistic (Edwards, 1972), the Akaike's information criterion (AIC), and the Bayesian information criterion (BIC).

As seen in Table 4, the lower the indices, the better the Rasch and bifactor models are fitted to the item response data. The results apparently show that the bifactor models are the best fitted models to the EFL item response data; however, the confirmatory bifactor models functioned better than the exploratory models in terms of goodness of fit. However, it is worth noting that the factor loading patterns in exploratory bifactor models are more interpretable than the confirmatory bifactor models. It is also shown that the Rasch TRT model is consistently better fitted to the EFL reading comprehension tests in comparison to the Rasch unidimensional IRT and MIRT models. Moreover, surprisingly, it revealed that the Rasch MIRT model does not necessarily better fit to the data in comparison to the conventional Rasch model. Therefore, in terms of the goodness of fit, the Rasch TRT, IRT, and the MIRT models are better fitted to the Iranian EFL reading comprehension data, respectively.

Discussion

The techniques of testing EFL reading comprehension are largely related to what we refer to reading comprehension. Reading comprehension is the ability to read and understand a given context quickly, which requires techniques of skimming and scanning the context, vocabulary recognition, comprehending questions, and giving correct grammatical and comprehensive responses to the related questions about the context (Henning, 1975). In doing so, almost in every occasions, passages/testlets are employed for testing reading comprehension. Testlet-based tests usually violate the assumption of unidimensionality that is required by the conventional IRT analysis because of the existing LID within the testlets, whereas the assumption is very difficult to be satisfied (Lee et al., 2001). TRT and bifactor models have dealt with the LID problem of the testlet reading passages and target the dimensionality assumption in several theoretical and empirical studies (e.g., Baghaei & Ravand, 2016; Dunn & McCray, 2020; Kim, 2017;

Table 4 Goodness of fit measures of applying the Rasch IRT, MIRT, and testlet models and the bifactor models on the EFL reading comprehension tests

Examination	Model	Goodness of fit measures		
		Log-likelihood	AIC	BIC
2016	Rasch IRT	− 16,288.60	32,599.21	32,668.98
	Rasch MIRT	− 16,380.62	32,785.23	32,861.35
	Rasch testlet	− 16,237.48	32,501.00	32,583.00
	Confirmatory Bifactor	− 16,098.53	32,257.00	32,447.00
	Exploratory Bifactor	− 16,112.51	32,283.00	32,467.00
2017	Rasch IRT	− 14,226.57	28,475.14	28,541.99
	Rasch MIRT	− 14,693.90	29,411.80	29,484.72
	Rasch testlet	− 14,206.54	28,439.00	28,518.00
	Confirmatory Bifactor	− 14,056.71	28,173.00	28,356.00
	Exploratory Bifactor	− 14,067.30	28,193.00	28,369.00
2018	Rasch IRT	− 22,116.68	44,255.37	44,325.90
	Rasch MIRT	− 22,936.52	45,897.05	45,973.99
	Rasch testlet	− 22,085.99	44,198.00	44,281.00
	Confirmatory Bifactor	− 21,880.97	43,822.00	44,014.00
	Exploratory Bifactor	− 21,881.98	43,822.00	44,008.00

Lee, 2004; Min & He, 2014; Morin et al., 2020; Ravand, 2015; Rijmen, 2010; Wang & Wilson, 2005; Wilson & Gochyyev, 2020). However, the magnitude of the testlet effect in the reading comprehension passages is still vague and it is not clear whether the effect is always non-negligible.

Moreover, due to popularity of the unidimensional and multidimensional Rasch and bifactor models in language testing, they have been extensively applied to the testlet-based reading comprehension passages (e.g., Aryadoust et al., 2020; Aryadoust & Zhang, 2016; Baghaei, 2012; Baghaei & Grotjahn, 2014; Baghaei & Carstensen, 2013; Boldt, 1992; Choi & Bachman, 1992; Jiao et al., 2013; Wang & Wilson, 2005). Nevertheless, the exploratory and confirmatory bifactor models and the Rasch testlet model have not yet been systematically compared to the unidimensional and multidimensional Rasch model through the real data of EFL reading comprehension testlets. Thus, the present research intended to fill the research gap by applying and comparing the bifactor and Rasch testlet model with the counterpart unidimensional and multidimensional Rasch models in the Iranian EFL testing context.

How large are the testlet effects for the EFL reading comprehension tests?

Addressing the first research question about the magnitude of the testlet effects in the EFL reading comprehension passages, the results showed that the testlet effects were all non-negligible. However, except for one testlet out of the all nine EFL reading comprehension passages, all of the testlet effects were small and non-substantial. DeMars (2012) believes that testlets might increase authenticity of the reading task as it adds more context to the test. However, modeling of negligible testlet effects makes the model unnecessary complicated and risks capitalization on chance, while it increases the error in parameter estimates. More precisely, ignoring the non-negligible testlet effect leads to overestimation of the classic reliability (Gessaroli & Folske, 2002; Li et al., 2010; Sireci et al., 1991; Wainer, 1995; Wainer and Wang, 2000), where at the same time, estimates of item discrimination parameters may be distorted (Wainer and Wang, 2000).

There is no accepted rule of thumb for evaluating the magnitudes of testlet effects, especially in the field of EFL language testing (Baghaei & Ravand, 2016). Nonetheless, Glas et al. (2000), Luo and Wolf (2019), Wang and Wilson (2005), Zhang (2010), and Zhang et al. (2010) considered testlet variances below 0.25 to be negligible, whereas testlet variances more than 0.50 supposed to be substantial by Luo and Wolf (2019), Wang et al. (2002, b), and Zhang (2010) in several empirical studies. Therefore, in terms of the magnitude of the testlet effects, most of the effects in the EFL reading comprehension tests were neither substantial nor negligible in this study. The results of this research are almost consistent with the findings of Wu and Wei (2014), as they claimed that the testlet effects were small but significant for the Chinese passage-based language tests. Nevertheless, the results contrast sharply with the findings of Kim (2017) in terms of the magnitudes of item dependencies within testlets.

DeMars (2012) considers testlet effect as a random nuisance factor which is not of interest in itself. She also believes that two general conceptions including LID and multidimensionality may be used to model the random nuisance factor and content specialist may waste time speculating on why some items load more than others within testlets. However, it is not still clear how random may be these effects and whether they are context-

dependent or not. Overall, considering the review of the research literature and the results of this study among the Iranian EFL students, it seems that the magnitudes of testlet effect are generally random and not dependent upon the context, at least in the field of reading comprehension testing in large-scale Iranian university entrance examinations.

Is there any correlation between item parameters (difficulty indices) of the Rasch testlet and bifactor models with each other and the other unidimensional and multidimensional Rasch models in the EFL reading comprehension tests?

Looking at the second research question, item parameters of the Rasch models were more consistent than the bifactor models in analyzing the EFL item response data, especially the Rasch testlet model had the most similar item difficulty parameters to the unidimensional Rasch model in comparison to the multidimensional Rasch model. That is, because most of the testlet effects were not substantial, the lower testlet effects yielded item difficulty parameters closer to the standard Rasch Model. If the testlet effects were zero, item parameters in the Rasch testlet model were exactly equal to the unidimensional Rasch model (Wang & Wilson, 2005). However, item difficulty parameters in the bifactor models were not correlated as perfect as the Rasch models, although the magnitudes of correlations were still high. This is probably because confirmatory bifactor model is a kind of restricted model in which item parameters are distorted in the process of estimation, whereas this problem does not occur in the exploratory version of bifactor analysis (Reise et al., 2010). Thus, lower magnitudes of correlations are expected between item difficulty parameters in the exploratory and confirmatory bifactor models in comparison to the Rasch models.

The results of Baghaei and Aryadoust (2015), Ravand (2015), and Wu and Wei (2014) are somehow in line with the results of this study, where they showed that item difficulty parameters were almost the same across IRT, TRT. However, the findings of this study contradict with the results of Min and He (2014), where they showed that item and ability parameters were the same in both TRT and bifactor models. Moreover, Wainer and Wang (2000) showed that the estimates of item difficulty parameters were not affected by LID, where testlet effects were mostly negligible within 50 EFL reading comprehension testlets. However, more research including simulation studies need to be done for further inference about the behavior of the parameters in different experimental conditions.

Are the Rasch testlet and bifactor models better fitted to the item response data of the EFL reading comprehension tests in comparison to the other unidimensional and multidimensional Rasch models?

At last, to answer the third research question, goodness of fit trials for the models showed that the bifactor models significantly fitted better than the other Rasch models to the EFL item response data. Bifactor models use marginal maximum likelihood (MML) estimation method to estimate item parameters, which permits conditional dependence within subset of items and provides more parsimonious factor solutions (Gibbons et al., 2007). On the other hand, Rasch testlet model also, as a special variation of the bifactor model (Li et al., 2006), outperformed in comparison to the other unidimensional and multidimensional Rasch models. Wang and Wilson (2005) showed

that item and person parameters as well as testlet effects could be more accurately recovered in Rasch testlet model in several experimental conditions. However, not all multidimensional models were fitted better than the unidimensional Rasch model in this study, because the multidimensional Rasch model was not fitted as well as the conventional Rasch model to the EFL test data. This result is in stark contrast to the findings of Baghaei (2012) and Baghaei and Grotjahn (2014), where they showed the superiority of the multidimensional models to assess the dimensionality of listening, reading, and the C-tests. LID is related to the dependence among the response functions of items within a testlet; however, this type of dependency cannot be captured by MIRT model (Andrich & Kreiner, 2010). Therefore, this may be a possible reason why MIRT model even not performed as well as the Rasch unidimensional model in terms of goodness of fit.

The results about the superiority of bifactor model fit against other IRT and TRT models confirm the findings of Byun and Lee (2016) and Foorman et al. (2015). However, the goodness of fit of the multidimensional model versus unidimensional Rasch model is in contradiction with the results of Baghaei (2012) and Baghaei and Grotjahn (2014). Nonetheless, The results about the goodness of fit of the Rasch testlet models were the same as the findings of Baghaei and Ravand (2016), Chang and Wang (2010), Eckes and Baghaei (2015), Huang and Wang (2013), Kim (2017), Rijmen (2010), and Wang and Wilson (2005).

In addition, confirmatory bifactor models fitted a little bit better than the exploratory bifactor models, but the factor loading patterns in the exploratory bifactor models were more interpretable than the confirmatory bifactor models for all of the EFL item response data. Exploratory analyses allow for directly analyzing the factor structures and identifying modeling problems, while they have no restrictions of confirmatory factor analysis methods to detect the problems after fitting models (Browne, 2001). Moreover, in exploratory bifactor analysis, the item and ability parameters may be distorted, because small cross-loadings are forced to zero and items with significant cross-loadings are accommodated on group factors (Finch, 2011). Reise (2012) insisted on inspecting the test dimensionality through an exploratory bifactor analysis prior to run the confirmatory model. Therefore, that is maybe why some researchers choose exploratory version of the bifactor model to explore the underlying factor structure of psychological measures (e.g., Reise, 2012; Reise et al., 2010). Altogether, the performance of the bifactor models looks very promising and evolving, whereas even some researchers are recently trying to blend the bifactor model with structural equation modeling methods (Morin et al., 2020).

Conclusion

The Iranian Ph.D. entrance examination is a national high-stakes test in Iran taken by more than one hundred thousand examinees every year, which makes it an influential test affecting a large number of master's degree holders hoping to pursue Ph.D. education at Iranian universities. In general, this research is the only study that opts for the application of the Rasch testlet model and bifactor analysis in the high-stakes Iranian university entrance examinations. On the one hand, this study was a successful experience of applying Rasch testlet and bifactor models on some high-profile EFL reading comprehension tests, and they did not suffer from some of the shortcomings and

limitations pointed out by studies deploying unidimensional IRT models to study EFL reading testlets (e.g., Choi & Bachman, 1992; Li et al., 2010). On the other hand, apart from identifying testlet effects underlying the EFL reading comprehension tests, it also benefited from newly introduced methods of assessing dimensionality. The findings, which were in many respects new to the field of language testing, might call for a re-evaluation of (the construct validity of) high-stakes examinations in light of more stringent alternative methodologies and models.

Although DeMars (2012) believes that the testlet effects are so random and erratic that she recommends subject matter specialists not to speculate on the sources of relationship between the factors and items; however, there is still a room for qualitative in-depth content analysis of EFL reading tests to find the different sources of LID. Quantitative research design can also be used to investigate for possible relationships between types of reading comprehension contents and the magnitudes of testlet effects among EFL learners.

Moreover, there are also promising measurement models, titled random block item response theory model (Lee & Smith, 2020) and composite model (Wilson & Gochyyev, 2020), which have been newly introduced to researchers and practitioners to deal with the LID and test dimensionality. Random block item response theory model is statistically equivalent to the Rasch testlet model, which allows for a more complicated and informative model including covariate variables such as gender and age into the model (Lee & Smith, 2020). The composite model consists of two parts simultaneously, including a multidimensional model for the subtests, and a predictive model for a composite of the latent variables based on each subtest. Composite model is claimed that has certain advantages over unidimensional, multidimensional, and hierarchical measurement models including TRT and the bifactor models (Wilson & Gochyyev, 2020). The aforementioned measurement models may be specifically considered for future application and comparison in handling testlet effects and test dimensionality analysis, especially in the field of language testing and assessment.

The limitation of the current study largely lies in the fact that the findings of this study are restricted to the limited number of EFL reading comprehension sections of the Iranian national university entrance examinations as the study populations. Hence, for more generalizability, future research may also adopt simulation studies (Morris et al., 2019) to investigate the prospect of Rasch testlet model and bifactor analysis in analyzing item response data. In addition, it is also recommended that the proposed measurement models, especially the bifactor models, be employed to investigate the dimensionality of other sections in high-stakes language testing in Iranian or other Asian EFL testing contexts. At last, I humbly encourage applied linguists and language testing practitioners/learners to benefit the implications and applications of Rasch testlet and bifactor models in testing reading.

Appendix

Technical formulas for the unidimensional, multidimensional, and the Rasch testlet models

The unidimensional version of the Rasch model (Rasch, 1960) is simply formulated as Formula 1:

$$p\left(y_{ij}=1\right)=\frac{\exp\left(\theta_i-b_j\right)}{1+\exp\left(\theta_i-b_j\right)} \tag{1}$$

where y_{ij} is the score of examinee i to the item j, $p(y_{ij}=1)$ is the probability that examinee i answer item j of a test correctly, and θ_i denotes the examinee ability and b_j is the item difficulty.

The multidimensional Rasch model is shown in Formula 2:

$$p\left(y_{ij}=1\right)=\frac{\exp\left(\theta_1+\theta_2+\ldots+\theta_k-b_j\right)}{1+\exp\left(\theta_1+\theta_2+\ldots+\theta_k-b_j\right)} \tag{2}$$

where θ_1, θ_2..., θ_k represent k latent traits or abilities of the examinee i, so in the model, the probability of answering an item correctly is function of more than one latent trait.

A three parameters testlet response theory (TRT) model is formulated as follows:

$$p\left(y_{ij}=1\right)=c_j+\left(1-c_j\right)\frac{\exp(a_j\left(\theta_i-b_j-\gamma_{id(j)}\right)}{1+\exp\left(a_j\left(\theta_i-b_j-\gamma_{id(j)}\right)\right)} \tag{3}$$

where a_j is the discrimination (slope) of the item j, c_j is the pseudo chance level parameter of the item j, and $\gamma_{id(j)}$ is the testlet parameter which is the effect of testlet *d(j)* on examinee i. Testlet effect actually captures the within testlet covariation in the TRT model. In this vein, the two parameters logistic (2PL) TRT and the Rasch testlet models are cases of the 3PL TRT model, when in 2PL TRT model $c_j=0$ and additionally the a_j is computed to be fixed in Rasch testlet model other than setting the c_j to zero (Wang & Wilson, 2005). Thus, in this context, the Rasch testlet model is reduced to the Formula (4) with the same aforementioned elements:

$$p\left(y_{ij}=1\right)=\frac{\exp\left(\theta_i-b_j-\gamma_{id(j)}\right)}{1+\exp\left(\theta_i-b_j-\gamma_{id(j)}\right)} \tag{4}$$

Moreover, considering the odd logarithm of answering over not answering an item correctly and doing a little algebra on the Eq. (2), the more simple Rasch testlet equation is obtained as

$$log\left(\frac{p\left(y_{ij}=1\right)}{p\left(y_{ij}=0\right)}\right)=\theta_i-b_j-\gamma_{id(j)} \tag{5}$$

If there is no testlet effect, $\gamma_{id(j)}=0$, the Rasch testlet model is equal to the standard Rasch model (Rasch, 1960).

Abbreviations

AIC: Akaike's information criterion; BIC: Bayesian information criterion; DIC: Deviance Information Criterion; EFL: English as a foreign language; ETS: Educational Testing Service; FCE: The First Certificate of English; IRT: Item response theory; GSEEE: Graduate School Entrance English Exam; IELTS: International English Language Testing System; LID: Local item dependency; MIRT: Multidimensional Item Response Theory; MKA: Minneapolis Kindergarten Assessment; MML: Marginal Maximum Likelihood; TOEFL: Test of English as Foreign Language; PIRLS: Progress in International Reading Literacy Study; TRT: Testlet response theory; 3PL: Three parameters logistic; 2PL: Two parameters logistic; UCLES: University of Cambridge Local Examinations Syndicate

Acknowledgements

The author would like to acknowledge the National Organization for Educational Testing of Iran for a grant for the data, without which the present study could not have been conducted.

Author's contributions
The author read and approved the final manuscript.

Author's information
Masoud Geramipour has a Ph.D. in Assessment and Measurement from the Faculty of Psychology and Education at Allameh Tabatabie University, Tehran. Currently, he is in the Department of Curriculum Studies and Educational Research at Kharazmi University, where he has taught research methodology and psychometrics courses to educational research students since 2010 as an assistant professor of educational research, measurement and statistics.

Funding
This work was supported by Kharazmi University.

Availability of data and materials
The data are available upon request from the author.

Competing interests
The author declares that he has no competing interests.

Received: 22 September 2020 Accepted: 18 January 2021
Published online: 19 February 2021

References

Ackerman, T. A. (1992). *Assessing construct validity using multidimensional item response theory*. San Francisco: *Paper presented at the Annual Meeting of American Educational Research Association* ERIC Document Reproduction Service No. ED344936.
Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.
Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, *38*(1), 123–140.
Andrich, D. (1988). *Rasch models for measurement (No. 68)*. London: Sage.
Andrich, D., & Kreiner, S. (2010). Quantifying response dependence between two dichotomous items using the Rasch model. *Applied Psychological Measurement*, *34*(3), 181–192.
Aryadoust, V., Ng, L. Y., & Sayama, H. (2020). A comprehensive review of Rasch measurement in language assessment: recommendations and guidelines for research. *Language Testing*, *0265532220927487*.
Aryadoust, V., & Zhang, L. (2016). Fitting the mixed Rasch model to a reading comprehension test: exploring individual difference profiles in L2 reading. *Language Testing*, *33*(4), 529–553.
Baghaei, P. (2012). The application of multidimensional Rasch models in large-scale assessment and validation: an empirical example. *Electornic Journal of Research in Educational Psychology*, *10*(1), 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*(1), 71–87.
Baghaei, P., & Carstensen, C. H. (2013). Fitting the mixed Rasch model to a reading comprehension test: identifying reader types. *Practical Assessment, Research and Evaluation*, *18*(1), 5.
Baghaei, P., & Grotjahn, R. (2014). Establishing the construct validity of conversational C-tests using a multidimensional Rasch model. *Psychological Test and Assessment Modeling*, *56*(1), 60.
Baghaei, P., & Ravand, H. (2016). Modeling local item dependence in cloze and reading comprehension test items using testlet response theory. *Psicológica*, *37*(1), 85–104.
Baldwin, S. G. (2007). A review of testlet response theory and its applications. *Journal of Educational and Behavioral Statistics*, *32*(3), 333–336.
Betts, J., Pickart, M., & Heistad, D. (2011). Investigating early literacy and numeracy: exploring the utility of the bifactor model. *School Psychology Quarterly*, *26*(2), 97.
Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, & M. R. Novick (Eds.), *Statistical theories of mental test scores (chapters 17-20)*. Reading: Addison-Wesley.
Boldt, R. F. (1992). Cross validation of item response curve models using TOEFL data. *Language Testing*, *9*(1), 79–95.
Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: fundamental measurement in the human sciences*, (3rd ed.,). New York: Routledge.
Briggs, D. C., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement*, *4*(1), 87–100.
Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*(1), 111–150.
Brunner, M., Nagy, G., & Wilhelm, O. (2012). A tutorial on hierarchically structured constructs. *Journal of Personality*, *80*(4), 796–846.
Byun, J. H., & Lee, Y. W. (2016). Investigating topic familiarity as source of testlet effect in reading tests: bifactor analysis. , *23*(3), 79–109.
Chalhoub-Deville, M., & Turner, C. E. (2000). What to look for in ESL admission tests: Cambridge certificate exams, IELTS, and TOEFL. *System*, *28*(4), 523–539.
Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, *48*(6), 1–29.
Chang, Y., & Wang, J. (2010). *Examining testlet effects on the PIRLS 2006 assessment*. Gothenburg: 4th IEA International Research Conference.
Choi, I. C., & Bachman, L. F. (1992). An investigation into the adequacy of three IRT models for data from two EFL reading tests. *Language Testing*, *9*(1), 51–78.
DeMars, C. E. (2012). *Confirming testlet effects*. *Applied Psychological Measurement*, *36*(2), 104–121.sa.

Drasgow, F., & Lissak, R. (1983). Modified parallel analysis: a procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied Psychology*, *68*, 363–373.
Dunn, K. J., & McCray, G. (2020). The place of the bifactor model in confirmatory factor analysis investigations into construct dimensionality in language testing. *Frontiers in Psychology*, *11*, 1357.
Eckes, T., & Baghaei, P. (2015). Using testlet response theory to examine local dependence in C-tests. *Applied Measurement in Education*, *28*(2), 85–98.
Edwards, A. W. F. (1972). *Likelihood*. Cambridge: Cambridge University Press.
Fan, J., Knoch, U., & Bond, T. (2019). Application of Rasch measurement theory in language assessment: using measurement to enhance language assessment research and practice. *Language Testing*, *8*(2), 3–9.
Finch, H. (2011). Multidimensional item response theory parameter estimation with nonsimple structure items. *Applied Psychological Measurement*, *35*(1), 67–82.
Foorman, B. R., Koon, S., Petscher, Y., Mitchell, A., & Truckenmiller, A. (2015). Examining general and specific factors in the dimensionality of oral language and reading in 4th–10th grades. *Journal of Educational Psychology*, *107*(3), 884.
Geramipour, M. (2020). Item-focused trees approach in differential item functioning (DIF) analysis: a case study of an EFL reading comprehension test. *Journal of Modern Research in English Language Studies*, *7*(2), 123–147.
Geramipour, M., & Shahmirzadi, N. (2018). Application and comparison of multidimensional latent class item response theory on clustering items in comprehension tests. *Journal of Asia TEFL*, *15*(2), 479.
Geramipour, M., & Shahmirzadi, N. (2019). A gender–related differential item functioning study of an English test. *Journal of Asia TEFL*, *16*(2), 674.
Gessaroli, M. E., & Folske, J. C. (2002). Generalizing the reliability of tests comprised of testlets. *International Journal of Testing*, *2*(3-4), 277–295.
Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., … Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement*, *31*(1), 4–19.
Glas, C. A. W., Wainer, H., & Bradlow, E. T. (2000). MML and EAP estimates for the testlet response model. In W. J. van der Linden, & C. A. W. Glas (Eds.), *Computer Adaptive Testing: Theory and Practice*, (pp. 271–287). Boston: Kluwer-Nijhoff Publishing.
Hambleton, R. K. (1989). *Principles and selected applications of item response theory*. In R. L. Linn (Ed.), *The American Council on Education/Macmillan series on higher education. Educational measurement*, (pp. 147–200). New York: Macmillan Publishing Co, Inc; American Council on Education.
Henning, G. (1992). Dimensionality and construct validity of language tests. *Language Testing*, *9*(1), 1–11.
Henning, G. H. (1975). Measuring foreign language reading comprehension. *Language Learning*, *25*(1), 109–114.
Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, *2*(1), 41–54.
Huang, H. Y., & Wang, W. C. (2013). Higher order testlet response models for hierarchical latent traits and testlet-based items. *Educational and Psychological Measurement*, *73*(3), 491–511.
Jiao, H., Kamata, A., Wang, S., & Jin, Y. (2012). A multilevel testlet model for dual local dependence. *Journal of Educational Measurement*, *49*(1), 82–100.
Jiao, H., Wang, S., & He, W. (2013). Estimation methods for one-parameter testlet models. *Journal of Educational Measurement*, *50*(2), 186–203.
Kim, W. H. (2017). *Application of the IRT and TRT models to a reading comprehension test* (Doctoral dissertation, Middle Tennessee State University).
Lee, G., Kolen, M. J., Frisbie, D. A., & Ankenmann, R. D. (2001). Comparison of dichotomous and polytomous item response models in equating scores from tests composed of testlets. *Applied Psychological Measurement*, *25*(4), 357–372.
Lee, H., & Smith, W. Z. (2020). A Bayesian random block item response theory model for forced-choice formats. *Educational and Psychological Measurement*, *80*(3), 578–603.
Lee, Y. W. (2004). Examining passage-related local item dependence (LID) and measurement construct using Q3 statistics in an EFL reading comprehension test. *Language Testing*, *21*(1), 74–100.
Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement*, *30*(1), 3–21.
Li, Y., Li, S., & Wang, L. (2010). Application of a general polytomous testlet model to the reading section of a large-scale English language assessment. *ETS Research Report Series*, *2010*(2), i–34.
Luo, Y., & Wolf, M. G. (2019). Item parameter recovery for the twoparameter testlet model with different estimation methods. *Psychological Test and Assessment Modeling*, *61*(1), 65–89.
Markon, K. E. (2019). Bifactor and hierarchical models: specification, inference, and interpretation. *Annual Review of Clinical Psychology*, *15*, 51–69.
McKinley, R. L., & Way, W. D. (1992). The feasibility of modeling secondary TOEFL ability dimensions using multidimensional IRT models. *ETS Research Report Series*, *1992*(1), i–22.
Mellenbergh, G. J. (2019). Test dimensionality. In *Counteracting Methodological Errors in Behavioral Research*, (pp. 135–156). Cham: Springer.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement*, (3rd ed., pp. 13–104). New York: NY American Council on education and Macmillan.
Min, S., & He, L. (2014). Applying unidimensional and multidimensional item response theory models in testlet-based reading assessment. *Language Testing*, *31*(4), 453–477.
Monseur, C., Baye, A., Lafontaine, D., & Quittre, V. (2011). PISA test format assessment and the local independence assumption. *IERI Monographs Series. Issues and Methodologies in Large-Scale assessments.*, *4*, 131–158.
Morin, A. J., Myers, N. D., & Lee, S. (2020). Modern factor analytic techniques: bifactor models, exploratory structural equation modeling (ESEM), and bifactor-ESEM. In *Handbook of sport psychology*, (pp. 1044–1073).
Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102.
Mundfrom, D. J., Shaw, D. G., & Ke, T. L. (2005). Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing*, *5*(2), 159–168.
O'Sullivan, B., & Council, B (2012). *Aptis test development approach. London: British Council*.

R Core Team (2019). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing URL https://www.R-project.org/.
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.
Ravand, H. (2015). Assessing testlet effect, impact, differential testlet, and item functioning using cross-classified multilevel measurement modeling. *SAGE Open*, *5*(2), 2158244015585607.
Reckase, M. D. (2009). Multidimensional item response theory models. In *Multidimensional item response theory*, (pp. 79–112). New York: Springer.
Reder, S. (1998). Dimensionality and construct validity of the NALS assessment. In *Literacy for the twenty-first century*, (pp. 37–57).
Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*(5), 667–696.
Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment*, *92*(6), 544–559.
Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement*, *47*(3), 361–372.
Rizopoulos, D. (2006). ltm: an R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, *17*(5), 1–25.
Robitzsch, A. (2019). *sirt: supplementary item response theory models. R package version 3*, (pp. 5–53) https://CRAN.R-project.org/package=sirt.
Robitzsch, A., Kiefer, T., & Wu, M. (2019). TAM: test analysis modules (R package Version 3.3-10). Verfügbar unter https://CRAN.R-project.org/package=TAM.
Rosenbaum, P. R. (1988). Items bundles. *Psychometrika*, *53*(3), 349–359.
Schindler, J., Richter, T., Isberner, M. B., Naumann, J., & Neeb, Y. (2018). Construct validity of a process-oriented test assessing syntactic skills in German primary schoolchildren. *Language Assessment Quarterly*, *15*(2), 183–203.
Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22*(1), 53–61.
Schwarz, G. (1978). The annals of statistics. *Estimating the Dimension of a Model*, *6*, 461–464.
Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*(3), 237–247.
Slocum-Gori, S. L., & Zumbo, B. D. (2011). Assessing the unidimensionality of psychological scales: Using multiple criteria from factor analysis. *Social Indicators Research*, *102*(3), 443–461.
Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*(4), 589–617.
Sun, J. (2005). Assessing goodness of fit in confirmatory factor analysis. Measurement and Evaluation in counseling and development 37(4), 256–240.
Tate, R. (2002). *Test dimensionality*. In G. Tindal, & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: validity, technical adequacy, and implementation*, (pp. 181–211). New Jersey: Lawrence Erlbaum Associates Publishers.
Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. *Applied Measurement in Education*, *8*(2), 157–186.
Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: an analog for the 3PL model useful in testlet-based adaptive testing. In *Computerized adaptive testing: theory and practice*, (pp. 245–269). Dordrecht: Springer.
Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York: Cambridge University Press.
Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement*, *37*(3), 203–220.
Wang, W.-C., Chen, P.-H., & Cheng, Y.-Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods*, *9*(1), 116–136.
Wang, W. C., Cheng, Y. Y., & Wilson, M. (2005). Local item dependence for items across tests connected by common stimuli. *Educational and Psychological Measurement*, *65*(1), 5–27.
Wang, W. C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, *29*(2), 126–149.
Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian Model for testlets: theory and applications. *Applied Psychological Measurement*, *26*(1), 109–128.
Westen, D., & Rosenthal, R. (2003). Quantifying construct validity: two simple measures. *Journal of Personality and Social Psychology*, *84*(3), 608.
Wilson, M. (1988). Detecting and interpreting local item dependence using a family of Rasch models. *Applied Psychological Measurement*, *12*(4), 353–364.
Wilson, M., & Gochyyev, P. (2020). Having your cake and eating it too: multiple dimensions and a composite. *Measurement*, *151*, 107247.
Wu, R., & Wei, J. (2014). Testlet effects on Chinese passage-based test. *China Examinations*, *12*, 9.
Yen, W. M. (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.
Zhang, B. (2010). Assessing the accuracy and consistency of language proficiency classification under competing measurement models. *Language Testing*, *27*, 119–140.
Zhang, O., Shen, L., & Cannady, M. (2010). *Polytomous IRT or testlet model: an evaluation of scoring models in small testlet size situations* (Unpublished Doctoral dissertation, University of Florida).

Publisher's Note