The Implementation of Testlet Models into Evaluating a Reading Comprehension Test

Do Thi Ha
HCMC University of Technology and Education, Vietnam

Abstract— Testlets, groups of test items that share the common input, are widely used in language testing. However, there lies a problem of Local Item Dependence (LID) which may result in adverse consequences for test score perception. This study analyzed the reading comprehension section of an English multiple-choice test to determine whether, and to what extent, testlet effects are inherent by using Testlet Response Theory (TRT). The data was gathered from 1653 non-English majors taking that final test at a university in Vietnam. The comparison of item difficulty and discrimination between Testlet Response model and Item Response Theory (IRT) model was made to illustrate the testlet effects. The paper, therefore, demonstrated how these models can be utilized into the test development process so that test developers can choose the best-fitting model and minimize inaccuracies of ability estimation.

Index Terms— Correlation matrix, Item discrimination, Item difficulty, Item response theory, Reading comprehension test, Testlet, Testlet response theory.

1 Introduction

Language tests are often known to consist of items with contents and traits intertwined. Especially when it comes to reading comprehension, there may be several related questions under a reading passage, which requires multiple skills to get through. Literature has recorded them as "item bundles" [1], [2], or "context-dependent item sets" [3], [4], [5], or "testlets" [6], among which the term "testlets" is mostly used by linguists and psychologists. Its utilization in educational assessment can be justified in two ways. First, students are taken on a tour of the text in which all of its different elements contribute to information decoding. Furthermore, item sets are beneficial in promoting higher levels of thinking such as analysis and synthesis than single items [7].

In spite of the above-mentioned engaging characteristics, educationalists need to take into consideration several stringent psychometric problems. The violation of local independence assumption, committed by the use of similar item types, shared subskills and passages, might lead to correlated responses to items within the same teslet. Therefore, what cannot be avoided is biased parameter estimation and test-equating errors [8]. A study by Thissen, Steinberg and Mooney [9] pointed out that such item correlation may affect test validity and result in Standard Error of Measurement (SEM). Chen and Thissen's [10] simulated data revealed that with LID, parameter estimates may differ from the case with locally independent data. Yen [11] also illustrated the effects of LID on students' trait measurement and biased Item Characteristic Curves (ICC), or in other words, on the test evaluation.

A lot of research has been done to build up compensation models for LID's deleterious effects. One of the most prevalent approaches to testlet models is the application of Multidimensional Item Response Theory (MIRT). Testlet Response Theory (TRT) models of Wainer, Bradlow and Wang [12] and Bifactor models of Gibbons and Hedeker [13] are both MIRT-oriented. Glas, Wainer and Bradlow's [14] study indicated that when there is a significant interrelation among test items, the IRT standard errors in estimating item difficulty and discrimination are greater than those of TRT. Drescher [15] also proved that regarding ability estimation, RMSE (Root Mean Square Error) of dichotomous IRT models is at a higher level. A research by Li, Bolt and Fu [16], moreover, showcased the equivalence between Bifactor and TRT models. This implies that in case of item covariation, Bifactor and TRT models stand out as a better choice than IRT.

In Vietnam, although testlets have long been embedded in all kinds of tests, they have yet to be of research interest, let alone TRT models. For all these reasons, it is vital that testlet effects be taken to a deeper level concerning test validity. The paper aims at investigating such effects on validating a Reading Comprehension test with TRT models as a measurement of item difficulty and discrimination.

2 Literature Review

2.1 Testlet

In 1987, Wainer and Kiely introduced the concept of testlet, "a group of items related to a single content area that is developed as a unit and contains a fixed number of predetermined paths that an examinee may follow"([6], p. 190). The classification of testlets is often based on item grouping. The first category entails items that have the same stimulus such as a reading comprehension test with an aggregation of items accompanying each passage. The second type refers to item sets that cover the same content area (e.g. subtests). Testlets have been extensively used in large-scale assessment, especially in the form of multiple-choice (MC) items. Some advantages of MC items can be listed as:

(i) Objectiveness: there is no worry about rater bias thanks to fixed answers [17]

• *Do Thi Ha is currently working as a lecturer at Ho Chi Minh City University of Technology and Education in Vietnam.*
Email: dothiha1985@gmail.com

(ii) Compared to constructed response items, MC items permit wide sampling and broad coverage of the content domain.

(iii) Consistency in marking and wide sampling promote test reliability and validity.

(iv) Simple scoring and efficient administering

Particularly, the assortment of MC items (i.e. testlets) helps measure different aspects of the cognitive activities. It incorporates a richer context into connected item sets, which makes test design and practice more flexible and effective. A wide range of learning outcomes, in this way, can be assessed through the interpretation of a single passage. According to DeMars [7], testlets, after all, bring authenticity to the test.

2.2 Local Item Dependence (LID)

One of the two crucial assumptions of IRT is Local Independence. If the test items are locally independent, students' responses to each item are not statistically related. However, testlets bring about complications into the theory and practice of educational measurement. Responses to items within a testlet have a tendency to be interrelated even after controlling for latent ability, which violates the assumption of conditional independence. Previous work by Thissen et al. [9] and Wainer [18] confirmed the latency of Local Item Dependence (LID) within testlets. Ferrara et al. [19], [20] and Yen [11] also found evidence for LID in reading comprehension tests. Thissen, Steinberg and Money [9] demonstrated that the item interconnection yielded assessment bias and wrong estimation of SEM. Using simulated data, Chen and Thissen [10] realized the difference in parameter estimates when comparing with the case of locally dependent data. A study by Yen [11] then pinpointed the negative effects of LID on students' construct evaluation and biased ICC, which leads to inaccurate estimation of the model's parameter. Other research [21], [22] corroborated the theory with the misinterpretation of parameter estimates when LID is neglected.

2.3 IRT Models

The preliminary ideas of IRT models were presented by Thustone [23], followed by Lord [24] with a notion of Item Characteristic Curves (ICC). In 1968, Birnbaum [25] applied logistic models for IRT, which were then built up by Lord and Novick [26], and then Bock and Aitkin [27]. Meanwhile, they developed some approaches to paramater estimation. One of the salient features of IRT is how it relates each examinee's latent traits to item parameter (i.e. item difficulty and discrimination) through his/her response to each question in the test [28], [29], [30]. Therefore, in IRT, ability parameters estimated are not test dependent and item parameters are sample independent [31].

In order to apply IRT, the following three assumptions need to be met:

- Unidimensionality: each test item measures a single latent trait.

- Monotonicity: the probability of each student's correct response will increase when his/her ability increases.

- Local independence: such probability is not affected by other examinees as well as the student's response to other items. IRT logistic model is presented as:

$$P\left(X_{ij}=1\,|\,\theta_j,a_i,b_i\right)=1+\exp\left(-a_i\left(\theta_j-b_i\right)\right) \tag{1}$$

in which, $\exp(\ )$ is exponential function with base e; $P\left(X_{ij}=1\,|\,\theta_j,a_i,b_i\right)$ denotes the probability of student j's correct response to item i; θ_j is student j's ability; a_i is item discrimination and b_i is its difficulty. According to Baker [30] and Hasmy [32], item discrimination and item difficulty can be classified respectively as in Tables 1 and 2:

Table 1
LABELS FOR ITEM DISCRIMINATION

Very High:	$a_i \geq 1.7$
High:	$1.35 \leq a_i < 1.7$
Moderate:	$0.65 \leq a_i < 1.35$
Low:	$0.35 \leq a_i < 0.65$
Very Low:	$a_i < 0.35$

Table 2
LABELS FOR ITEM DIFFICULTY

Very Hard:	$b_i \geq 2$
Hard:	$0.5 \leq b_i < 2$
Medium:	$-0.5 \leq b_i < 0.5$
Easy:	$-2 \leq b_i < -0.5$
Very Easy:	$b_i < -2$

Besides, IRT models can be expressed by normal ogive function with the use of integral Gaussian function.

$$P\left(X_{ij}=1\,|\,\theta_j,a_i,b_i\right)=\int_{-\infty}^{a_i\left(\theta_j-b_i\right)}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt \tag{2}$$

In comparison with the former, the latter ICC are slightly steeper for the same set of item parameter values. To make up for this difference, Birnbaum [34] recommended the multiplication of the exponents in the logistic model by 1.7 to make the two models more similar. The nomarl ogive function in logistic model will then have the form as follows:

$$P\left(X_{ij}=1\,|\,\theta_j,a_i,b_i\right)=1+\exp\left(-1.7a_i\left(\theta_j-b_i\right)\right) \tag{3}$$

2.4 TRT Models

In 2005, Wang and Wilson [33] put forward Rasch testlet model with the addition of another different dimension for testlet effects. Their model can be represented in the function below:

$$P_{jdi}=\frac{\exp\left(\theta_j-b_i+\gamma_{jd_i}\right)}{1+\exp\left(\theta_j-b_i+\gamma_{jd_i}\right)} \tag{4}$$

Later, Wainer et al. [12], [34] extended Rasch model into two and three-parameter logistic models by adding discrimination parameter a_i and guessing parameter c_i. When given in normal ogive function, their TRT model can be formulated as:

$$P_{jd_i}=1+\exp\left(-1.7a_i\left(\theta_j-b_i+\gamma_{jd_i}\right)\right) \tag{5}$$

where P_{jd_i} is the probability of item i's correct response in

testlest d_i. Without testlet effects (i.e. $\gamma jd_i = 0$), the model turns into standard two-parameter IRT model. Testlet-based local item dependence manifests itself through the testlet effect variance $\sigma^2_{jd_i}$. That is, the greater the testlet effect variance of Testlet d_i is, the higher the degree of associated local item dependence is [35], [36]. If the testlet effect variance is zero, there is no indication of local dependence within the testlet [34].

However, some still cast doubt on the relativity and objectiveness of the testlet variance. Glas, Wainer and Bradlow [14] reckoned that for simulation studies, the variances below 0.25 can generally be considered negligibly small. Meanwhile, Zhang [37] indicated that in empirical studies of a language test, the variance ranges from 0.5 to 2 and higher. Another approach by Min and He [38] applied χ^2 tests for each pair of test items, which may pose a drawback if there are considerable testlets and items within them.

Taking the above-mentioned research as guidelines, the author adapted testlet models for validating a reading comprehension test for non-English majors, which so far has not been investigated statistically and appropriately.

3 Objective and Methodology

3.1 Research Objective

The purpose of the study is to examine testlet effects inherent in the English multiple-choice test. Therefore, the following questions were addressed for test evaluation:

- Whether, and to what extent, does Local Item Dependence (LID) exist in the Reading Comprehension (RC) section of the test? If so, for what item sets are testlet models suitable?

- How can item difficulty and discrimination be estimated using testlet models? Is there any significant difference in measurement results when TRT and IRT 2PL models are used? And what is the better-fitting model?

3.2 Instruments

The data for this study was gathered randomly from 1653 students taking the English final test at a university in Ho Chi Minh city, Vietnam. Each correct answer is coded 1; 0 is allocated to the other cases (including incorrect, null or inappropriate choices).

The test consists of three sections, among which Reading Comprehension (RC) ranges from item 31 to item 60. According to Wainer and Kielly [6], RC items are better treated as testlet data to control the local dependence. When comparing Testlet 2PL with IRT and MIRT models for RC test of GSEEE (Graduate School Entrance English Exam), Min and He [37] reached the same conclusion that Testlet models provide more appropriate analyses for RC tests. That justifies why in this study, only 30 multiple-choice RC items (items 31-60) were investigated for students' intended abilities. Henceforth, these 30 items will be referred to as the test for more convenience.

R is a free software used for statistical computing in recent research because of its flexibility [39], [40]. The pakage distributed by R can be easily downloaded free of charge at http://CRAN.R-project.org.

3.3 Methodology

The RC section consists of 9 reading texts which are equivalent to 9 testlets [6]. Firstly, Bartlett's test was conducted to compare the observed correlation matrix to the identity matrix [41]. If the values outside the main diagonal are often high (in absolute value), some variables are correlated; if most of these values are close to zero, the PCA (Principal Component Analysis) is not really useful. Under Ho, $|R| = 1$; if the variables are highly correlated, we have $|R| \neq 0$.

The Bartlett's test statistics indicate to what extent we deviate from the reference situation $|R| = 1$. After that, TRT 2PL and IRT 2PL models were employed to measure item difficulty and discrimination. The package psych dealt with investigating LID, and the package sirt was used for parameter estimation of TRT and IRT 2PL models. The comparison of the results helps decide the better-fitting model for such data.

4 Data Analysis

4.1 Local Item Dependence

Based on Wainer and Kielly [6], the RC test was divided into 9 testlets associated with 9 common stimuli (i.e. reading passages) as follows:

- Testlet 1: Items 31-33.
- Testlet 2: Items 34-38.
- Testlet 3: Items 39-41.
- Testlet 4: Items 42-44.
- Testlet 5: Items 45-47.
- Testlet 6: Items 48-50.
- Testlet 7: Items 51-53.
- Testlet 8: Items 54-56.
- Testlet 9: Items 57-60.

Bartlett's Sphericity Test was applied to investigate the existence of LID in the testlets.

Table 3
BARTLETT'S SPHERICITY TEST

	Chi square	Degree of freedom	p-value
Testlet 1	79.318	3	***
Testlet 2	118.043	10	***
Testlet 3	227.11	3	***
Testlet 4	483.259	3	***
Testlet 5	27.564	3	***
Testlet 6	21.907	3	***
Testlet 7	20.556	3	**
Testlet 8	136.059	3	***
Testlet 9	129.895	6	***
			*: p-value<0.05
			**: p-value<0.001
			***: p-value<0.0001

Table 3 shows statistical significance of the Bartlett's test in 9 testlets, or put it another way, there is a presence of LID among the testlets. Such results reinforce the notions of

Wainer and Kiely [6] as well as Min and He [38].

4.2 Item Difficulty and Discrimination

The package sirt of freeware R was exploited to examine item difficulty and discrimination in both TRT and IRT 2PL models. The results can be found in Table 4:

Table 4
ITEM DIFFICULTY & DISCRIMINATION
OF TRT MODEL

		TRT	
		a	**b***
Testlet 1	Item31	0.248	1.96371
	Item32	0.334	1.67964
	Item33	2.896	-0.2842
Testlet 2	Item34	0.303	2.70627
	Item35	0.349	0.53582
	Item36	0.261	2.88123
	Item37	0.477	1.10273
	Item38	0.655	0.36641
Testlet 3	Item39	1.505	-1.5176
	Item40	0.634	-1.1577
	Item41	0.839	-2.0799
Testlet 4	Item42	3.124	0.15973
	Item43	1.297	0.79241
	Item44	0.122	11.9918
Testlet 5	Item45	0.163	2.63804
	Item46	0.14	3.37857
	Item47	3.667	-0.0693
Tetslet 6	Item48	0.056	10.9821
	Item49	0.264	-0.4545
	Item50	1.358	0.65243
Testlet 7	Item51	4.779	-0.023
	Item52	0.025	23.84
	Item53	0.167	-0.976
Testlet 8	Item54	0.499	0.3006
	Item55	0.516	0.25775
	Item56	0.947	0.25238
Testlet 9	Item57	0.35	-1.7514
	Item58	2.591	0.85218
	Item59	0.405	1.04444
	Item60	-0.012	-21.083

In Table 4, the values in Column a represent item discrimination and b* item difficulty. Based on Baker [30] and Hasmy [32]'s labels for item difficulty and discrimination (as shown in Tables 1 and 2), it can be deduced that:

- Item difficulty (b*): 2 items are categorized as Very Easy; 4 items as Easy, and 9 Medium items, all of which add up to 30%. 8 items are ranked as Hard and 7 Very Hard, which account for 50% of the test. With 50% of items at Hard and above levels, the RC test is supposed to be challenging. In particular, at an extremely low level of -21.083, Item 60 does need revision. The reason may lie in the fact that this question could be answered based on mere Common Sense.

- Item discrimination (a): over 40% (13 items) have Very Low discrimination. 6 items (20%) are at Low level. 3 items fit in Moderate level. There are 3 questions at High and 5 Very High levels. With a majority of items (60%) at bad discrimination levels, it can be concluded that although the test is difficult for examinees' ability, it fails to distinguish between students with higher and lower levels of knowledge. Further analysis navigated my concern to some test items. Items 48, 52 and 60 should be redesigned for greater discriminating power. As mentioned earlier, for some items, students do not need to read the text for their answers. In other words, they do not really test examinees' reading comprehension abilities.

In addition, the results expose a contradictory issue in psychometrics. The most demanding questions turn out to be at the lowest discrimination level. That may result from a decent amount of guessing behavior when test-takers encounter such questions.

Table 5 below illustrates mean comparison of item difficulty and discrimination between TRT and IRT models.

Table 5
MEAN COMPARISON
OF ITEM DIFFICULTY & DISCRIMINATION

	p-value of t-test		DIC	
	a	**b***	**TRT**	**IRT**
Testlet1	0.44032	0.57959	5524.57	5433.5
Testlet2	0.73283	0.21978	9855.89	9885.91
Testlet3	0.54432	0.39943	3588.9	3467.02
Testlet4	0.8084	0.71819	3619.78	3688.33
Testlet5	0.43326	0.02054	3594.59	6273.4
Testlet6	0.64621	0.23389	6018.01	6033.34
Testlet7	0.45121	0.42429	5223.09	6201.24
Testlet8	0.67538	0.38544	6475.85	6471.01
Testlet9	0.46444	0.38843	7428.94	7717.67

The mean comparison acknowledges no considerable distinction between TRT and IRT models, except for the difficulty level of Testlet 5. Therefore, DIC got involved to measure how well the models fit the data. According to Spiegelhaler, Best, Carlin and Van der Linde [42], TRT model with smaller DIC (Deviance Information Criterion) is regarded as more suitable for the dataset.

5 Discussion and Implication

This study investigated the application of testlet models to validate a reading comprehension test. A 30-item excerpt of an English multiple-choice test was taken into consideration when TRT is the best choice. The results reflect testlet effects inherent in a language test as in any kind of assessment, which reinforces Wainer and Kiely [6] as well as Min and He [38]'s ideas of item correlation.

Generally, the test is supposed to be tough for the test takers with 50% of items at difficult level, but have bad discriminating power (over 60% of items at low level). More often than not, high-challenge questions tend to distinguish well among student. Nonetheless, there are cases in which difficult items have mediocre discrimination, which can be justified by guessing parameter. Taken a closer look, Items 34, 36, 45, 46, 48, 52 and 54 may involve high values of guessing parameter. This implies that students' guesswork, rather than know-

ledge, may engage in figuring out the answers.

To accommodate LID, Eckes [35] and Ravand [36] used testlet effect variance while Min and He [38] exploited χ^2 test for each pair of items in each testlet. As mentioned earlier, these two methods still have some disadvantages, i.e. the relativity of testlet effect variance and the complexity of χ^2 tests. Therefore, this study's use of Bartlett's test for LID examination is somewhat more practical and convincing. Item parameter estimation was conducted using Markov chain Monter Carlo (MCMC) methods. DIC utilization is a different approach to the model suitability from Min and He [38] and Eckes [35].

Further research will focus on comparing TRT with IRT, MIRT and Bifactor models. Besides, contradictory findings on difficult items with bad discrimination have prompted an application of extended TRT models for proper estimation. More qualitative item analyses will also be needed to determine how well they meet the learning goals and what may contribute to students' reading trouble.

REFERENCES

[1] P.R. Rosenbaum, "Item bundles", *Psychometrika,* vol. 53, pp. 349-359, 1988.

[2] M. Wilson and R.J. Adams, "Rasch model for item bundels", *Psychometrika,* vol. 60, pp. 181-198, 1985.

[3] T.M. Haladyna, "Context-dependent item sets", *Educational Measurement: Issues and Practice,* vol. 11, no. 1, pp. 21-25, 1992.

[4] T.M. Haladyna, *Developping and validating multiple-choice test items.* (3rd ed), Mahwah, NJ: Lawrence Elrbaum, 2004.

[5] L.A. Keller, H. Swaminathan and S.G. Sireci, "Evaluating scoring procedures for context-dependent item sets", *Applied Measurement in Education*, vol. 16, pp. 207-222, 2003.

[6] H. Wainer and G.L. Kiely, "Item clusters and computerized adaptive testing: A case for testlet", *Journal of Educational Measurement*, vol. 24, pp. 185-202, 1987.

[7] C.E. DeMars, "Application of the bi-factor multidimensional item response theory model to testlet-based test", *Journal of Educationa Measurement*, vol. 43, pp. 145-168, 2006.

[8] J. Chen, "Model selection for IRT equating of testlet-based tests in the random group design", unpublished doctoral dissertation, University of Iowa, Iowa City, 2014.

[9] D. Thissen, L. Steinberg and J.A. Mooney, "Trace lines for testlets: A use of multiple-catogorical-reponse model", *Journal of Education Measurement*, vol. 26, pp. 247-260, 1989.

[10] W.H. Chen and D. Thissen, "Local dependence indexes for item pairing using item response theory", *Journal of Educational and Behavioral Statistics*, vol. 22, pp. 265-289, 1997.

[11] W.M. Yen, "Scaling performance assessments: Strategies for managing local item dependence", *Journal of Educational Measurement*, vol. 30, pp. 187-213, 1993.

[12] H. Wainer, E.T. Bradlow and X. Wang, *Testlet response theory and its application,* Cambridge: Cambridge University Press, 2007.

[13] R.D. Gibbons D.R. Hedeker, "Full-information bi-factor analysis", *Psychomatrika*, vol. 57, pp. 423-436, 1992.

[14] C.A.W. Glas, H. Wainer and E.T. Bradlow, "MML and EAP estimation in testlet-based adapting test", In W.J van der Linder and C.A.W. Glas, *Computerized adapting test: Theory and practice*, Boston, MA: Kluwer-Nijhoff, pp. 271-288, 2000.

[15] A.R. Drescher, "An emperical investigation of LID using the testlet model: A further look", Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA, 2004.

[16] Y. Li, D.M. Bolt and J. Fu, "A comparison of alternative models for testlets", *Applied Psychological Measurement*, vol. 30, pp. 3-21, 2006.

[17] J. Millman and Greene, "The specification and development of tests of achievement and ability", In R.L. Linn, *Educational Measurement*, 3rd Ed, New York: American Council on Education and Macmillan, pp. 335-366, 1989.

[18] H. Wainer, "Precision and differential item functioning on a teslet-based test: The 1991 Law School Admission Test as an example", *Applied Measurement in Education*, vol. 8, pp. 157-186, 1995.

[19] S. Ferarra, H. Huynh and H. Baghi, "Contexttual characteristics of locally dependent open-ended item clusters in a large scale performace assessment", *Applied Measurement in Education*, vol. 10, pp. 123-144, 1997.

[20] S. Fererra, H. Huynh and H. Michaels, "Contextual explanation of local dependence in item clusters in a large scale hands-on science performance assessment", *Journal of Educational Measurement*, vol. 36, pp. 119-140, 1999.

[21] L.M. Reese, *The impact of local dependencies on some LSAT outcomes*, Report No. LSAC-R-95-02, Newton, PA: Law School Admission Council, ERIC Document Reproduction Service No. ED469244, 1995.

[22] I.H.-S. Ip, "Adjusting for information inflation due to local dependency in moderately large item clusters", *Psychometrika*, vol. 65, pp. 73-91, 2000.

[23] L.L. Thurstone, "A method of scaling psychological and education test", *Journal of Education Psychology*, vol. 16, pp. 433-451, 1925.

[24] F.M. Lord, "A theory of test score", *Pychometric Monographs*, vol. 7, Richmon, VA: Psychometric Coporation, Retrieved from http://www.psychometrika.org/journal/online/MN07.pdf, 1952.

[25] A. Birnbaum, "Some latent trait models and their use in infering an examinees' ability", In F.M. Lord and M.R. Novick, *Statistical theories of mental test scores*, Reading, MA: Addition-Wesley, pp. 395-479, 1968.

[26] F.M. Lord and M.R. Novick, *Statistical theory of mental test score*, Reading, MA: Addition-Wesley, 1968.

[27] R.D. Bock and M. Aitkin, "Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm", *Psychometrika*, vol. 46, no. 4, pp. 433-459, 1981.

[28] B.D. Wright and M.H. Stone, *Best test design*, Chicago: MESSA Press, 1979.

[29] G. Camilli and L.A. Shepard, *Methods for identifying biased test items*, vol. 4, Thousand Oaks, CA: Sage, 1994.

[30] F.B. Baker, *The basic of item response theory,* USA: ERIC Clearinghouse on Assessment and Evaluation, 2001.

[31] R.K. Hambleton and H. Swaminathan, *Item response theory: Principles and applications*, USA: Kluwer-Nijhoff, 1985.

[32] A. Hasmy, "Compare unidimensional and multidimensional Rasch model for test with multidimensional construct and item local dependence", *Journal of Education and Learning*, vol. 8, no. 3, pp. 187-194, 2014.

[33] W.-C Wang and M.R. Wilson, "The Rasch testlet model", *Applied Psychological Measurement,* vol. 29, pp. 126-149, 2005.

[34] H. Wainer and X. Wang, "Using a new statistical model for testlet to

score TOEFL", *Journal of Educational Measurement,* vol. 37, pp. 203-220, 2000.

[35] T. Eckes, "Examining testlet effects in the TestDaF listening section: A testlet response theory modeling approach", *Language Testing,* vol. 31, no. 1, pp. 39-61, 2014.

[36] H. Ravand, "Assessing testlet effect, impact, defferential testlet and item functioning using cross-classified multilevel measurement modeling", *SAGE Open*, April-June, pp. 1-9, 2015.

[37] B. Zhang, "Assessing the accuracy and consistency of language proficiency classification under competing measurement models", *Language Testing,* vol. 27, pp. 119-140, 2010.

[38] S. Min and L. He, "Applying unidimensional and multidimensional item response theory model in testlet-based reading assessment", *Language Testing*, vol. 31, no. 4, pp. 453-477, 2014.

[39] K. Kelley, K. Lai and P.J. Wu, "Using R for data analysis: A best practice for research", In J.W. Osborne, *Best advanced practices in quantitative methods*, Thousand Oaks, CA: Sage, pp. 535-572, 2008.

[40] A. Vance, "Data analysis captivated by R's power", New York Times, Retrieved from http://www.nytimes.com/2009/01/07/technology/busine%20ss-computing/07program.html/?pagewanted=all, 2009.

[41] M.S. Bartlett, "The effect of standardization on a χ^2 approximation in factor analysis", *Biometrika*, vol. 38, no. 3, pp. 337-344, 1951.

[42] D.J. Spiegelhalter, N.G. Best, B.P. Carlin and A. van der Linder, "Bayesian measures of model complexity and fit (with discussion)", *Journal of Royal Statistical Society*, Ser B, vol. 64, pp. 583-639, 2002.