International Journal of Assessment Tools in Education
2022, Vol. 9, No. 1, 254–266

https://doi.org/10.21449/ijate.948182

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate **Research Article**

Item parameter recovery via traditional 2PL, Testlet and Bi-factor models for Testlet-Based tests

Sumeyra Soysal[1,*], Esin Yilmaz Kogar[2]

[1]Necmettin Erbakan University, Ahmet Keleşoğlu Faculty of Education, Department of Educational Sciences, Educational Measurement and Evaluation, Konya, Turkiye
[2]Niğde Ömer Halisdemir University, Faculty of Education, Department of Educational Sciences, Educational, Measurement and Evaluation, Niğde, Turkiye

ARTICLE HISTORY

Received: June 04, 2021
Revised: Jan. 22, 2022
Accepted: Feb. 03, 2022

Keywords:
Local dependence,
Testlet model,
Bi-factor model,
Random effect parameter,
Parameter recovery.

Abstract: The testlet comprises a set of items based on a common stimulus. When the testlet is used in the tests, there may violate the local independence assumption, and in this case, it would not be appropriate to use traditional item response theory models in the tests in which the testlet is included. When the testlet is discussed, one of the most frequently used models is the testlet response theory (TRT) model. In addition, the bi-factor model and traditional 2PL models are also used for testlet-based tests. This study aims to examine the item parameters estimated by these three calibration models of the data properties produced under different conditions and to compare the performances of the models. For this purpose, data were generated under three conditions: sample size (500, 1000, and 2000), testlet variance (.25, .50, and 1), and testlet size (4 and 10). For each simulation condition, the number of items in the test was fixed at $i = 40$ and 100 replications were made under each condition. Among these models, it was concluded that the TRT model gave less biased results than the other two models, but the results of the bi-factor model and the TRT were more similar as the sample size increased. Among the examined conditions, it was determined that the most effective variable in parameter recovery was the sample size.

1. INTRODUCTION

Item response theory (IRT, Lord & Novick, 1968) is a model that is widely used for test development and test scoring, because of its strong mathematical modeling. One of the important assumptions of this theory is local independence (LI). This assumption is generally expressed as "the examinee's trait (ability or proficiency) value, denoted provides all the necessary information about the examinee's performance, and once trait level is considered, all other factors affecting examinee performance are random" (Wainer et al., 2000, p.248). However, this assumption can be violated when the items share a common stimulus. In the literature, such items are generally referred to as testlets.

The concept of testlet, first expressed by Wainer and Kiely (1987), is the name given to a group of items associated with a single comprehensive stimulus. The testlet shows a set of items that

[*]CONTACT: Sumeyra Soysal ✉ sumeyrasoysal@hotmail.com Necmettin Erbakan University, Ahmet Keleşoğlu Faculty of Education, Department of Educational Sciences, Educational Measurement and Evaluation, Konya, Turkiye

share a single common stimulus, such as a reading passage or an information graph, and where performance on each item depends on both a general ability and a specific ability related to a specific content or situation (Li, 2017). Such items are widely used in many national and international large-scale tests because of their various advantages. For example, testlets allow over one item to be asked based on the same stimulus, allowing over one information to be collected from a stimulus, thus improving the efficiency of the test (information per unit time) (Wainer et al., 2000). Another advantage of testlets is that they help develop a more realistic and context-based test (Li, 2017). Through these context-dependent items, measuring higher-level skills may become more workable (DeMars, 2006). It is known that testlets provide a significant advantage in computer adaptive test (CAT) applications. In CAT applications, there is a specific item selection algorithm for each person. Here, there is a context-effect caused by the content of the items. However, this effect is reduced, as individuals will encounter the same context when they take the same testlet. In short, the use of testlets in CAT applications provides greater control of the negative effects of single items, allowing as much fairness as possible among test takers (Pak, 2017). However, in such items, some students have a special interest or better prior background knowledge in a passage than other students, in this situation they are likely to perform better on the items related to this passage than on other items of the same difficulty level, or they tend to perform better than other students with the same general ability level (Li, 2017, p.1). Therefore, testlets lead to the emergence of additional sources of variance, such as content knowledge (Chen & Thissen, 1997). DeMars (2006) states that responses to items in a testlet may be related to testlet-specific background knowledge or skills, or to a secondary characteristic, such as testlet-specific interest or other motivational factors. This situation has revealed the necessity of a special examination of testlet items.

Another disadvantage of testlets is that testlets violate the LI assumption of the unidimensional IRT. Although this assumption is violated, the use of unidimensional IRT models for such items leads to inaccurate in parameter estimations (Sireci et al., 1991; Wainer & Wang, 2000; Yen, 1993). Therefore, different models have been developed to handle testlets. The psychometric framework that deals with testlets is known as testlet response theory (TRT) models (Bradlow et al., 1999; Wainer et al., 2000; Wainer et al., 2007). This model includes one more parameter explaining the interaction between each item and each examinee within a testlet, besides the parameters in the traditional IRT model for dichotomous items. Another solution to model the dependency among test items in testlets is the bi-factor model (Rijmen, 2010). Recently, multilevel models have also been used to address local item dependence among items (Jiao et al., 2005; Jiao et al., 2012). These models consider local item dependence because of item clustering. In addition, although there are testlets, it is very common to apply the traditional IRT model to items that are scored dichotomously. Because when the testlet effect is determined to be low, the negligibility of this effect or the usability of traditional IRT estimations, which are more familiar to researchers, are discussed in the literature (Glas et al., 2000; Eckes, 2014; Eckes & Baghaei, 2015; Min & He, 2014). It has also been examined with polytohomus IRT models that treat testlets as a single item, and it has been stated that there is a need for models that give more information about testlets (Wainer, 1995).

Since unidimensional, testlet and bi-factor models are widely used in testlet examinations, it is important to evaluate whether the parameters estimated from these models can be accurately estimated. Since all conditions cannot be tested on the real data set, this study was carried out on simulation data. The advantage of knowing the real parameter values in simulation studies makes the accuracy of the estimation method is measurable. This study, it is aimed to examine the data produced under different testlet conditions with traditional two-parameter logistic (2PL), the TRT, and the bi-factor models by varying the sample size, the size of the testlet variance, and the number of testlets. Koizol (2016) stated that the bi-factor model did not receive enough attention in testlet reviews. Liu and Liu (2012) stated that it is not clear to practitioners in which cases traditional IRT models should be used instead of a newly proposed testlet model. In this study, it is aimed to provide more helpful information to practitioners by considering many possible conditions. It is expected that this study, which also includes the bi-factor model in testlet examinations, will contribute to filling the gaps in this subject. Because of this study, determining the conditions under which local item dependency has serious effects on parameter estimations with the help of many conditions tested can guide the researchers in choosing the right model.

1.1. Calibration Models

There are strategies developed over different models to deal with the local independence assumption violation caused by testlets. Traditionally, the items in the testlets have been treated as independent items like other items in the test, and traditional IRT models have been used as the calibration model. The traditional 3PL model for dichotomous data is specified as

$$P_{ni}\,(1) = c_i + \,(1 - \,c_i)\,\frac{\exp[a_i\,(\theta_n - \,b_i\,)]}{1 + \exp[a_i\,(\theta_n - \,b_i\,)]}, \tag{1}$$

where $P_{ni}(1)$ is the probability of response 1 (correct) for person n on the ith item; a_i, b_i, and c_i are the discrimination, difficulty, and guessing parameters, respectively; and θ_n is person's ability. However, this approach has been found to cause biased parameter estimation and overestimation of test reliability (Sireci et al., 1991; Thissen et al., 1989; Tuerlinckx & De Boeck, 2001; Wainer & Thissen, 1996; Wainer & Wang, 2000; Yen, 1993).

In another approach, the testlets were evaluated as a single item and scored in polytomous. Although this approach has been found to yield partially good results (Wainer, 1995), it has several shortcomings. For example, this approach is insufficient for situations that require more information about the items in the tests. Because of this approach, the testlet score is represented by the sum of the correct number (Wainer et al., 2000). These total scores lose answer pattern knowledge for each individual test taker. This loss of information can lead to an increase in measurement errors, which directly reduces overall test reliability (Keller et al., 2003; Sireci et al., 1991; Yen, 1993; Zenisky et al., 2002). Since polytomous models were not used in this study, the details of the model were not included.

Although it is stated that the violation of local independence does not cause serious problems when the length of the testlets is moderate (4-6 items/testlet), it is stated that as the testlets get longer, a special psychometric model is needed that can control local dependence (Wainer et al., 2007). In addition, in these models, attention should be paid to maintaining the item level as the unit of analysis. Bradlow et al. (1999) proposed a TRT model by adding a parameter (a testlet effect parameter) to the traditional 2PL model for items nested in the same testlet. This parameter represents the dependence between items within the same testlet, and the variances of the random testlet effects were assumed to be constant across testlets (Wang & Wilson, 2005). Later, this model was developed for 3PL (Wainer et al., 2000). The model is

$$P_{ni}\,(1) = c_i + \,(1 - \,c_i)\,\frac{\exp\left[a_i\left(\theta_n - \,b_i\, - \,\gamma_{nd(i)}\right)\right]}{1 + \exp\left[a_i\left(\theta_n - \,b_i\, - \,\gamma_{nd(i)}\right)\right]}, \tag{2}$$

As seen, in this model, unlike Equation 1, there is a $\gamma_{nd(i)}$ parameter. $\gamma_{nd(i)}$ is the random effect for person n on testlet $d_{(i)}$, which describes the interaction between persons and items within the testlet. The testlet effect is a random effect variance caused by local item dependency (LID), and the greater the variance, the greater the effect in the testlet (Wainer & Wang, 2000). The size of the variance shows the size of the dependence between the items. When the testlet random effect is zero, this model is equal to the traditional 3PL model.

The advantages of this model compared to polytomous models are expressed as follows (Wang et al., 2000; Wang & Wilson, 2005). The unit of analysis is the test items, not the testlets, so that the information in the response patterns within the testlets is not lost. The second advantage is that familiar item parameter concepts, such as item discrimination and item difficulty, are still valid and functional. Another advantage is that the standard item scoring scales (1 for a correct answer and 0 for an incorrect answer) remain unchanged. Thus, an easy transfer from the traditional IRT model to using the TRT model is provided.

This model uses the same item discrimination for both the theta and testlet traits, and this has been discussed as a limitation of the model (Li et al., 2004). When the data do not fit this constraint, the model is misspecified (Koziol, 2016). To handle discrimination parameters for these traits separately, the responses to the testlet items can be handled within the bi-factor model, which is a multidimensional model. It is already known that the testlet model is a special case of the bi-factor model (Rijmen, 2010). The answers given to the items in the bi-factor model are a function of both the primary trait and one of the secondary traits, and when this model is considered in the context of the testlet, secondary traits are testlet traits (DeMars, 2006). In this model, unlike the 3PL testlet model, the discrimination of an item on theta is not constrained proportionally to the discrimination on the corresponding testlet trait. The bi-factor model for dichotomous data is

$$P_{ni}\,(1) = c_i + \,(1 - \,c_i)\,\frac{\exp(a_{ip}\Theta_{np} + a_{is}\Theta_{ns} + d_i\,)}{1 + \exp(a_{ip}\Theta_{np} + a_{is}\Theta_{ns} + d_i\,)}, \tag{3}$$

where a_{ip} is the ith item slope parameter for the primary trait, a_{is} is the ith item slope for the sth secondary trait, θ_{np} is the nth person latent trait score for the primary dimension, θ_{ns} is the nth person latent score for the sth secondary trait, d_i is the ith item intercept parameter (d_i = -$a_i b_{\mathrm{i}}$), and c_i is the ith item guessing parameter. So, the TRT model in equation 2 can be viewed as a special case of the more general the bi-factor model in equation 3.

The testlet effect was investigated by simulation studies under different conditions. In these studies, different estimation methods improved item estimations (Luo & Wolf, 2012), equating methods were examined (Tao & Cao, 2016), evaluation of model comparison criteria (DeMars, 2012), ability parameter estimations were improved in CAT applications (Pak, 2017). The focus is on cases such as examining the effects when there are the different number of response categories (Wang et al., 2002). There are also studies in the literature evaluating parameter estimations got from different models with a similar purpose to the current study (Bradlow et al., 1999; DeMars, 2006; Koziol, 2016). The difference of this research from the mentioned studies is that it deals with more simulation conditions together.

2. METHOD

2.1. Simulation Design

Three independent variables were manipulated: a) sample size: 500, 1000, and 2000; b) testlet number: 40 dichotomous items in 4 or 10 testlets (10 items per each of 4 testlets and 4 items per each of 10 testlets); c) variance of the testlet effect: .25, .50, and 1, representing small to large effects. Wang and Wilson (2005, p.133) stated that the variances of the testlets in the real tests can be very diverse (from as small as almost zero to as large as the variance of the latent trait). In this study, the latent trait was generated with a standard normal distribution [θ ~ N(0, 1)]. Therefore, the largest variance of the testlet was chosen as 1.00. In this study, total 18 simulation conditions are considered, since a fully crossed design is used. Number of items was fixed to 40 to mimic a test of relatively medium length.

2.2. Data Generation

Similar to DeMars (2012), item discrimination and difficulty parameters were generated from a log-normal distribution N(0,1) ranging from .5 to 2.0 and a standard normal distribution N(0,1), respectively. Ability parameter and testlet variance were also generated from N(0,1) for the three possible testlet variance values determined by the specific simulation condition (same as Luo and Wolf, 2019, p.71). Based on these specifications, 40 dichotomously scored item response data were randomly generated. 100 replications were implemented for 18 conditions. Data generation was carried out through the R program.

2.3. Data Analysis

Each simulated data set for the traditional IRT, the TRT, and the bi-factor and model was calibrated using the *mirt* package (Chalmers, 2020) in R programme with the full information with the maximum likelihood estimation method with expectation-maximization (EM) algorithm. The stopping rule of the EM algorithm was set to the number of iterations = 500 or when maximum change = .00010. The models mentioned in the calibration model title in the previous section are as 3PL models. However, the guessing parameter was not considered in this study and the 2PL versions of the models were used. Because the three-parameter TRT model may encounter the problem of model convergence in practice (Wainer et al., 2007).

The performance of the three models is assessed using four criteria: the root-mean-square-error (RMSE) (i.e., total error), the bias (i.e., systematic error), mean absolute error (MAE), and the correlation between the estimated parameters and the true parameters. They are defined as;

$$RMSE\ (\hat{\pi}) = \sqrt{\frac{\sum_1^N \sum_1^R (\hat{\pi}_r - \pi)^2}{R\ X\ N}}\ , \tag{4}$$

$$Bias\ (\hat{\pi}) = \frac{\sum_1^N \sum_1^R (\hat{\pi}_r - \pi)}{R\ X\ N}, \tag{5}$$

$$MAE\ (\hat{\pi}) = \frac{\sum_1^N \sum_1^R |(\hat{\pi}_r - \pi)|}{R\ X\ N}, \tag{6}$$

where $\hat{\pi}_r$ is the estimated model parameter for the rth replication, π is the true model parameter, R is the number of replications, and N is the number of items.

3. FINDINGS

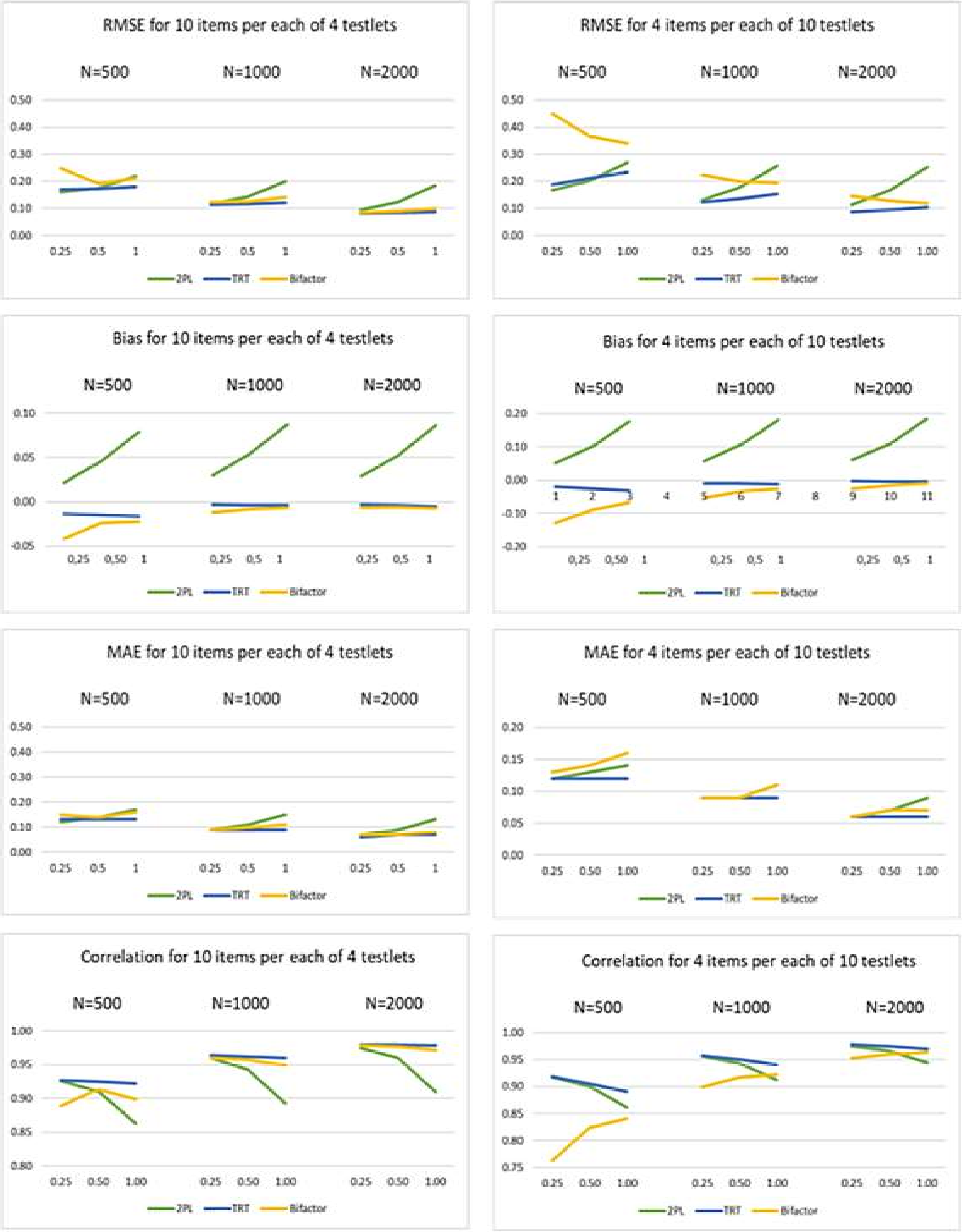
The recovery of item discrimination and item difficulty parameters across calibration models and testlet size conditions are presented in Figure 1 and Figure 2, respectively. Also, the complete set of results are summarized in the appendix as Table A1 and Table A2, respectively.

3.1. Recovery of Item Discrimination Parameters

As seen in Figure 1, under all conditions, the TRT model outperformed the traditional 2PL and bi-factor models concerning to the RMSE, the bias, the MAE of the estimated item discriminations, and correlations between the estimated and true item discriminations. For 10 items per each of 4 testlets, the performance of the TRT model outperformed with increased sample size but nearly remained stable across testlet variance (which RMSES were .17, .12 and .08 across sample size 500, 1000, and 2000, respectively). With same pattern, MAEs were .13, .9, and .7 across sample size 500, 1000, and 2000, respectively. The bi-factor model showed better recovery with increased sample size, but its performance slightly decreased with increased testlet variance.

The bi-factor model performed equivalently to the TRT model when the sample size was especially 1000 and 2000, which differences of RMSE and of correlation between the estimated and true item discriminations never exceeded .01 and MAE never exceed .02. This model showed the worst recovery under N = 500 condition when the criterion was correlation. Overall, the traditional 2PL model was the worst, showing large number of non-convergence conditions with increased testlet variance compared to both the TRT model and the bi-factor model. This means that EM cycles terminated after 500 iterations, not when the maximum change = .00010.

Figure 1. *Recovery of item discrimination.*

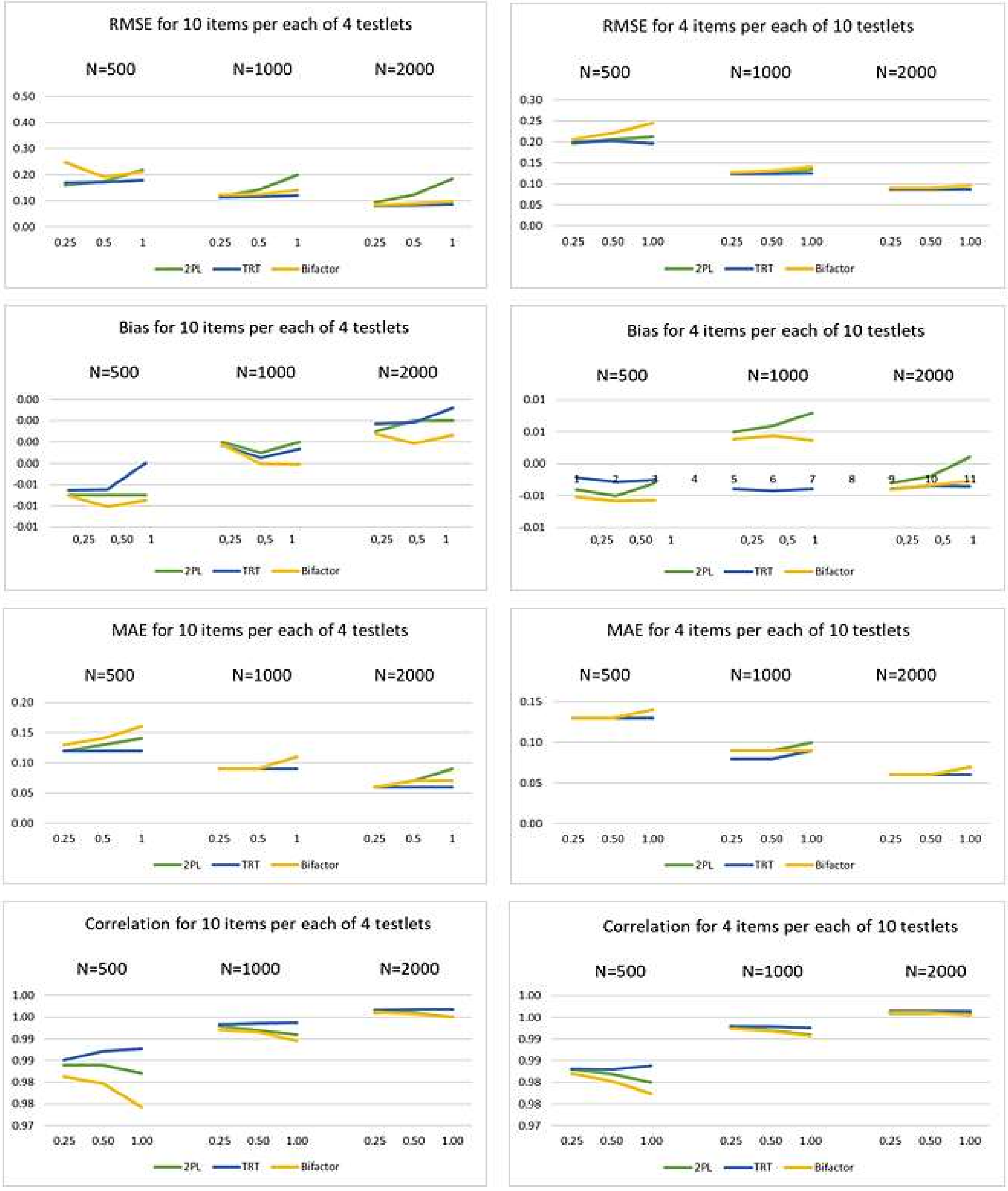


The bias showed that under all conditions, the traditional 2PL model underestimated the discrimination parameters than the true ones but the opposite tendency for the TRT and bi-factor models. For 4 items per each of 10 testlets, a similar pattern was held for the three calibration models with worse recovery, and the TRT model outperformed the other calibration models. The largest difference in RMSE between 10 items per each of 4 testlets and 4 items per each of 10 testlets was .07, .05, and .20 for the traditional 2PL, the TRT, and the bi-factor models, respectively. The bi-factor model showed more non-convergence conditions, especially when the sample size was 500. According to the correlation, the performance of the bi-factor model decreased with compared to the 10 items per each of 4 testlets, although the performance of the other two models nearly remained stable (which the differences never exceed .01).

3.2. Recovery of Item Difficulty Parameters

As seen in Figure 2, under all conditions, the TRT model slightly outperformed the traditional 2PL and bi-factor models with respect to the RMSE, the bias, the MAE of the estimated item difficulties, and correlations between the estimated and true item difficulties.

Figure 2. *Recovery of item difficulty.*



When the outcome criteria was RMSE for 10 items per each of 4 testlets, the performance of the TRT model performed better with increased sample size and slightly better with increased testlet variance under N = 500 but remained stable across testlet variance when the sample size was 1000 and 2000. Besides, for the MAE under all sample sizes, the TRT model performed stable across testlet variance. The bi-factor model outperformed recovery with increased sample size, but its performance slightly decreased with increased testlet variance. Considered RMSE, this model showed the worst performance under the N = 500 condition. The traditional 2PL model had the same pattern as the bi-factor model when the criteria were MAE under both 10 items per each of 4 testlets and 4 items per each of 10 testlets conditions. Again, the magnitude of bias ranged from .00 to .01, and correlations between the estimated and true item difficulty ranged from .98 to 1.00 and were the same across three calibration models and testlet size conditions. The differences in RMSE and in MAE were quite small, the largest difference between 10 items per each of 4 testlets and 4 items per each of 10 testlets was .02, .02, and .03 for the traditional 2PL, the TRT, and the bi-factor models, respectively.

4. DISCUSSION and CONCLUSION

Using testlets in tests violates the LI assumption. The TRT model and the bi-factor model have been widely used by researchers and practitioners to address local item dependency among the items in the same testlet. Besides these models, traditional 2PL models continue to be used for tests with testlets. In this study, dichotomous data simulated under different conditions (sample size, testlet size, and testlet variance size) were handled with three calibration models, the traditional 2PL, the TRT, and the bi-factor models, and the performances of the item parameters got from these three models were compared.

The TRT model outperformed the traditional 2PL and the bi-factor models regarding testlet size conditions, types of parameters, and outcome criteria. When the sample size was small, the performance of the bi-factor model was the worst under all other conditions and showed an irregular pattern. The reason is why a few item parameters in several replications were estimated quite differently from the true values, insomuch that RMSE was even .80 within in the replication itself. Besides, such a situation was not encountered in small samples for MAE, which produced more regular results. In this study, the stopping rule of the EM algorithm was set to the number of iterations = 500 or when maximum change = .00010. In all conditions where N = 500 and in some conditions for N = 1000, the EM cycles in the bi-factor model estimations stopped when they reached the maximum iteration. This had been an attempt to increase errors of the model estimation a little more than the normal. For the TRT model, a similar situation was observed in far less replication for N = 500. The time of the TRT and the bi-factor model estimations got longer under conditions of the large number of testlet, but the estimation time for the traditional 2PL model was barely or never impacted.

For both the traditional 2PL model and especially the bi-factor model, discrimination parameter recovery accuracy was negatively affected by increased testlet variance and the number of testlet but almost remained stable in the TRT model. The three calibration models themselves performed similar difficulty parameter recovery under conditions of the small number of items per testlet and the large number of items per testlet. Increased number of sample size was positively affected by both two types of parameter recovery for the three calibration models, especially the traditional 2PL and the bi-factor models. These findings are in line with the findings of DeMars (2006), Liu and Liu (2012), and Koziol (2016), who generated the data according to the TRT model (as was done in our study), that the performance of the TRT model was the best to the traditional 2PL model and the bi-factor model. Koziol (2016) examined the recovery of the parameters under only sample size was 1000 and used MAE to compare the efficacy of the three calibration models for recovery of item and person parameters. Our findings on recovery of the item discrimination and parameters with MAE (in Appendix, Table

A1 and Table A2) under N = 1000 were highly consistent with Koziol (2016). In contrast, Koziol (2016) reported that recovery of the item difficulty parameter only suffered under the largest testlet dependency condition (i.e., the large testlet variance and the large number of items per testlet condition). The difference between the current study and Koziol's findings (2016) could arise out of the estimation methods used within these two studies.

Sample size had a bigger impact on item parameter estimates than the other testlet conditions. Because the data followed the TRT model in this study, item parameters recovered the best with this model, as expected. In case of fully crossing the data generation according to calibration models in additional research, recovery and accuracy of parameters can be examined. However, under a large sample size and a small number of testlet, the performance of the bi-factor model could be as good as the TRT model. Also, under small testlet variance for any sample size, performing the traditional 2PL model could be as good as the TRT model. It should not be forgotten that even minor differences can have significant consequences in high-stakes contexts. Therefore, it is considered that more studies are needed on the parameter recovery and accuracy of modeling approaches. As with all studies, the results based on this study are limited to the conditions (i.e., testlet variance, the number of items per each of testlet, sample size, calibration models) given by the method. In this study, only the recovery of the item parameters was examined, the recovery of the ability parameter could be examined to vary outcome criteria and testlet conditions for future research. Also, another limitation of the present study is that we only used a medium-length test. The size of the number of items in the test can also be considered as a condition of the study.

Although testlet item structures are used in large-scale testing applications or classroom assessment, testlet dependency is generally ignored when calculating the test scores of individuals. As in this study, the effect of testlet dependency may be small or insignificant, but we do not know the exact magnitude of this effect in real-world testing situations. Therefore, as Koziol (2016) pointed out, it needs to be investigated whether test results will be biased if the testlet dependency is neglected or modeled incorrectly. To conclude, the findings of the current study show that the testlet and the bi-factor models provide to handle with LID and these two models give similar results in large samples.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors.

Authorship Contribution Statement

Sumeyra Soysal: Investigation, Methodology, Simulation Study, Formal Analysis, and Writing-original draft. **Esin Yilmaz Kogar**: Investigation, Resources, Methodology, Formal Analysis, and Writing-original draft.

Orcid

Sumeyra Soysal https://orcid.org/0000-0002-7304-1722
Esin Yilmaz Kogar https://orcid.org/0000-0001-6755-9018

REFERENCES

Bradlow, E.T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*(2), 153-168. https://doi.org/10.1007/bf02294533

Chalmers, R.P. (2020). *mirt: Multidimensional item response theory*. R package version 1.33.2. [Computer software manual]. http://www.R-project.org/

Chen, W.H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265–289. https://doi.org/10.3102/10769986022003265

DeMars, C.E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement, 43*(2), 145-168. https://doi.org/10.1111/j.1745-3984.2006.00010.x

DeMars, C.E. (2012). Confirming testlet effects. *Applied Psychological Measurement, 36*, 104-121. https://doi.org/10.1177/0146621612437403

Eckes, T. (2014). Examining testlet effects in the TestDaF listening section: A testlet response theory modeling approach. *Language Testing, 31*(1), 39-61. https://doi.org/10.1177/0265532213492969

Eckes, T., & Baghaei, P. (2015). Using testlet response theory to examine local dependence in C-tests. *Applied Measurement in Education, 28*(2), 85-98. https://doi.org/10.1080/08957347.2014.1002919

Glas, C.A.W., Wainer, H., & Bradlow, E.T. (2000). MML and EAP estimation in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 271–288). Kluwer-Nijhoff.

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. https://doi.org/10.1111/j.1745-3984.2011.00161.x

Jiao, H., Wang, S., & Kamata, A. (2005). Modeling local item dependence with the hierarchical generalized linear model. *Journal of Applied Measurement, 6*(3), 311-321.

Keller, L., Swaminathan, H., & Sireci, S.G. (2003). Evaluating scoring procedures for context-dependent item sets. *Applied Measurement in Education, 16*, 207-222. https://doi.org/10.1207/s15324818ame1603_3

Koziol, N.A. (2016). Parameter recovery and classification accuracy under conditions of testlet dependency: a comparison of the traditional 2PL, testlet, and bi-factor models. *Applied Measurement in Education, 29*(3), 184-195. https://doi.org/10.1080/08957347.2016.1171767

Li, F. (2017). *An information-correction method for testlet-based test analysis: From the perspectives of item response theory and generalizability theory* (Report No. ETS RR-17-27). ETS Research Report Series. https://doi.org/10.1002/ets2.12151

Liu Y, & Liu H.Y. (2012). When should we use testlet model? A comparison study of Bayesian testlet random-effects model and standard 2-pl bayesian model. *Acta Psychologica Sinica, 44*(2), 263-275. https://doi.org/10.3724/sp.j.1041.2012.00263

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Luo, Y., & Wolf, M.G. (2019). Item parameter recovery for the two-parameter testlet model with different estimation methods. *Psychological Test and Assessment Modeling, 61*(1), 65-89.

Min, S., & He, L. (2014). Applying unidimensional and multidimensional item response theory models in testlet-based reading assessment. *Language Testing, 31*(4), 453-477. https://doi.org/10.1177/0265532214527277

Paek, I., & Cole, K. (2019). *Using R for item response theory model applications*. Routledge.

Pak, S. (2017). *Ability parameter recovery of a computerized adaptive test based on rasch testlet models* [Doctoral dissertation, University of Iowa]. Iowa University Libraries https://doi.org/10.17077/etd.5akqn3gy

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*, 361-372. https://doi.org/10.1111/j.1745-3984.2010.00118.x

Sireci, S.G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*(3), 237-247. https://doi.org/10.1002/j.2333-8504.1991.tb01389.x

Tao, W., & Cao, Y. (2016). An extension of IRT-based equating to the dichotomous testlet response theory model. *Applied Measurement in Education, 29*(2), 108-121. https://doi.org/10.1080/08957347.2016.1138956

Thissen, D., Steinberg, L., & Mooney, J.A. (1989). Trace lines for testlets: A use of multiplecategorical-response models. *Journal of Educational Measurement, 26*, 247-260. https://doi.org/10.1111/j.1745-3984.1989.tb00331.x

Tuerlinckx, F., & De Boeck, P. (2001). The effect of ignoring item iterations on the estimated discrimination parameters in item response theory. *Psychological Methods, 6*(2), 181-195. https://doi.org/10.1037/1082-989x.6.2.181

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*, 157–186. https://doi.org/10.1207/s15324818ame0802_4

Wainer, H., Bradlow, E.T., & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In W. J. van der Linden & C.A.W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 245–270). Kluwer-Nijhoff.

Wainer, H., Bradlow, E.T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge University Press.

Wainer, H., & Kiely, G.L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185-201. https://doi.org/10.1111/j.1745-3984.1987.tb00274.x

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 15*(1), 22-29. https://doi.org/10.1002/j.2333-8504.1998.tb01749.x

Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement, 37*(3), 203-220. https://doi.org/10.1002/j.2333-8504.2001.tb01851.x

Yen, W.M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187-213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x

Zenisky, A.L., Hambleton, R.K., & Sired, S.G (2002). Identification and evaluation of local item dependencies in the Medical College Admissions Test. *Journal of Educational Measurement, 39*(4), 291-309. https://doi.org/10.1111/j.1745-3984.2002.tb01144.x

APPENDIX

Table A1. *Recovery of item discrimination parameters.*

Calibration Model	Conditions		Testlet Size							
			10 items per each of 4 testlets				4 items per each of 10 testlets			
	SS	TV	RMSE	Bias	MAE	Corr	RMSE	Bias	MAE	Corr
Traditional 2PL	500	.25	.16	.02	.12	.93	.17	.05	.13	.92
		.50	.17	.05	.14	.91	.20	.10	.15	.90
		1.00	.22	.08	.17	.86	.27	.18	.21	.86
	1000	.25	.12	.03	.09	.96	.13	.06	.10	.96
		.50	.14	.05	.11	.94	.18	.11	.14	.94
		1.00	.20	.09	.15	.89	.26	.18	.20	.91
	2000	.25	.09	.03	.07	.98	.11	.06	.09	.98
		.50	.12	.05	.09	.96	.17	.11	.13	.97
		1.00	.18	.09	.13	.91	.25	.18	.19	.94
Testlet Response Theory	500	.25	.17	-.01	.13	.93	.19	-.02	.14	.92
		.50	.17	-.01	.13	.92	.21	-.03	.15	.91
		1.00	.18	-.02	.13	.92	.23	-.03	.17	.89
	1000	.25	.11	.00	.09	.96	.12	-.01	.09	.96
		.50	.12	.00	.09	.96	.14	-.01	.10	.95
		1.00	.12	.00	.09	.96	.15	-.01	.11	.94
	2000	.25	.08	.00	.06	.98	.09	.00	.07	.98
		.50	.08	.00	.07	.98	.09	.00	.07	.97
		1.00	.09	.00	.07	.98	.10	.00	.08	.97
Bi-factor	500	.25	.25	-.04	.15	.89	.45	-.13	.24	.76
		.50	.19	-.02	.14	.91	.37	-.09	.21	.82
		1.00	.21	-.02	.16	.90	.34	-.07	.20	.84
	1000	.25	.12	-.01	.09	.96	.22	-.05	.13	.90
		.50	.13	-.01	.10	.96	.20	-.03	.13	.92
		1.00	.14	-.01	.11	.95	.19	-.03	.12	.92
	2000	.25	.08	-.01	.07	.98	.14	-.02	.09	.95
		.50	.09	-.01	.07	.98	.13	-.02	.08	.96
		1.00	.10	-.01	.08	.97	.12	-.01	.08	.96

Note. RMSE: Root mean square error, MAE: Mean absolute error, Corr: Pearson correlation coefficient.

Table A2. *Recovery of item difficulty parameters.*

Calibration Model	Conditions		Testlet Size							
			10 items per each of 4 testlets				4 items per each of 10 testlets			
	SS	TV	RMSE	Bias	MAE	Corr	RMSE	Bias	MAE	Corr
Traditional 2PL	500	.25	.19	-.01	.12	.98	.20	.00	.13	.98
		.50	.19	-.01	.13	.98	.21	-.01	.13	.98
		1.00	.20	-.01	.14	.98	.21	.00	.14	.98
	1000	.25	.13	.00	.09	.99	.13	.01	.09	.99
		.50	.13	.00	.09	.99	.13	.01	.09	.99
		1.00	.15	.00	.11	.99	.14	.01	.10	.99
	2000	.25	.09	.00	.06	1.00	.09	.00	.06	1.00
		.50	.10	.00	.07	1.00	.09	.00	.06	1.00
		1.00	.12	.00	.09	1.00	.10	.00	.07	1.00
Testlet Response Theory	500	.25	.19	-.01	.12	.99	.20	.00	.13	.98
		.50	.18	-.01	.12	.99	.20	.00	.13	.98
		1.00	.17	.00	.12	.99	.20	.00	.13	.98
	1000	.25	.12	.00	.09	.99	.12	.00	.08	.99
		.50	.12	.00	.09	.99	.12	.00	.08	.99
		1.00	.12	.00	.09	.99	.13	.00	.09	.99
	2000	.25	.08	.00	.06	1.00	.09	.00	.06	1.00
		.50	.08	.00	.06	1.00	.09	.00	.06	1.00
		1.00	.08	.00	.06	1.00	.09	.00	.06	1.00
Bi-factor	500	.25	.22	-.01	.13	.98	.21	-.01	.13	.98
		.50	.23	-.01	.14	.98	.22	-.01	.13	.98
		1.00	.27	-.01	.16	.97	.24	-.01	.14	.98
	1000	.25	.14	.00	.09	.99	.13	.00	.09	.99
		.50	.14	.00	.09	.99	.13	.00	.09	.99
		1.00	.16	.00	.11	.99	.14	.00	.09	.99
	2000	.25	.09	.00	.06	1.00	.09	.00	.06	1.00
		.50	.10	.00	.07	1.00	.09	.00	.06	1.00
		1.00	.11	.00	.07	1.00	.10	.00	.07	1.00

Note. RMSE: Root mean square error, MAE: Mean absolute error, Corr: Pearson correlation coefficient.