Psychological Test and Assessment Modeling, Volume 61, 2019 (1), 65-89

# Item parameter recovery for the two-parameter testlet model with different estimation methods

*Yong Luo*[1] *& Melissa Gordon Wolf*[2]

**Abstract**

A simulation study was conducted to investigate how MCMC, MMLE, and WLSMV, all implemented in Mplus, recovered item parameters and testlet variance parameter of the two-parameter logistic (2PL) testlet model. The manipulated factors included sample size and testlet variance magnitude, and parameter recoveries were evaluated with bias, standard error, root mean square error, and relative bias. We found that there were no statistically significant differences regarding parameter recovery between the three estimation methods. When both sample size and magnitude of testlet variance were small, both WLSMV and MCMC had convergence issues, which did not occur to MCMC regardless of sample size and testlet variance. A real dataset from a high-stakes test was used to demonstrate the estimation of the 2PL testlet model with the three methods.

Keywords: testlet model, estimation, MCMC, MMLE, WLSMV

[1] *Correspondence concerning this article should be addressed to:* Yong Luo, Phd, National Center for Assessment, West Palm Neighborhood, King Khalid Road, Riyadh 11534, Saudi Arabia.; email: jacky-luoyong@gmail.com

[2] University of California, Santa Barbara, California, USA

## Introduction

Item response theory (IRT; Lord, 1980) models refer to a family of statistical models that describe how the interaction between latent traits and item characteristics drives item responses. As statistical models, IRT models have assumptions and in order to fully reap the theoretical benefits of IRT such as the desirable property of parameter invariance, those assumptions must not be violated. One pivotal assumption of IRT is local item independence (LII), which stipulates that an examinee's responses to any pair of items should be independent after conditioning on the latent variable of interest. The assumption of LII, however, is often too stringent to be tenable in practical situations.

In educational assessments, the LII assumption is often violated when a group of items share a common stimulus, such as a reading comprehension passage or a mathematical diagram. For example, items belonging to the same reading comprehension passage tend to be correlated with each other even after conditioning on the latent construct. Such a phenomenon is known as local item dependence (LID; Yen, 1994) in the psychometric literature. LID results in the violation of LII and failure to address LID can lead to serious psychometric consequences, such as biased estimation of item parameters (Ackerman, 1987), overestimation of test reliability (Sireci, Thissen, & Wainer, 1991), equating bias (Tao & Cao, 2016), and erroneous classification of examinees into incorrect performance categories (Zhang, 2010). As the psychometric consequences of LID are grave, various methods have been proposed to detect and address LID (e.g., Bradlow, Wainer, & Wang, 1999; Braeken, Tuerlinckx, & De Boeck, 2007; Hoskens & De Boeck, 1997; Ip, 2002; Wilson & Adams, 1995).

The testlet model (Wainer, Bradlow, & Wang, 2007) is arguably the most popular approach to addressing LID. Due to its widespread usage in the psychometric community, it is important to evaluate whether the model parameters of a testlet model can be accurately estimated. Given that there are several estimation methods available, it is imperative to investigate which method leads to the most accurate estimation and provide researchers and practitioners with advice regarding which methods work best under various data conditions.

In this article, we compare the performances of three estimation methods: Markov chain Monte Carlo (MCMC), marginal maximum likelihood estimation (MMLE; Bock & Aitkin, 1981), and weighted least square adjusted by mean and variance (WLSMV; Muthén, du Toit, & Spisic, 1997) for the two-parameter-logistic (2PL) testlet model. We utilized Mplus (Muthén & Muthén, 1998-2012) as the software program because it provides a wide variety of estimation methods. Among the three methods, MCMC (the BAYES estimator) and MMLE (the MLR estimator) implemented in Mplus are both full information methods, in that they use the full multi-way frequency table (or in the IRT terminology, the whole response vector). Conversely, WLSMV operates as a limited information method in that it uses only the two-way frequency table. It should be noted that MCMC itself does not necessarily entail a full information method (for an illustration of using MCMC as a limited information method, see, Bolt, 2005). In addition, when a MMLE estimator is used, Mplus employs the full-information bi-factor analysis meth-

od (Gibbons & Hedecker, 1992) to simplify the multidimensional integration issue to two-dimensional integration and hence reduce the computation burden.

The remainder of the articles is structured as follows. First, we explain why we focused on the 2PL testlet model instead of the more general three-parameter-logistic (3PL) testlet model, followed by a demonstration of the 2PL testlet model as a special case of the bi-factor model, and a brief discussion of the three estimation methods investigated in the present study. Second, we provide a literature review on simulation studies that investigate parameter estimation of the testlet model, excluding its multilevel and higher-order extensions. Third, we present a simulation study conducted to compare the item parameter recovery of the 2PL testlet model with the three estimation methods. Fourth, we use a real dataset to demonstrate the estimation of the 2PL testlet model with these methods. Finally, we conclude the current paper with discussions and recommendations regarding the parameter estimation of the 2PL testlet model in practice.

## Background and literature review

### Estimation difficulty of the 3PL testlet model

While the 3PL multidimensional IRT (MIRT; Reckase, 2009) model is more generalized than its 2PL counterpart, the latter is often preferred in practice. Such preferences can be attributed to the fact that the pseudo-guessing parameter in the 3PL MIRT models is usually difficult to estimate, and researchers (e.g., McLeod, Swygert, & Thissen, 2001; Stone & Yeh, 2006) often resort to a two-step procedure, in which the pseudo-guessing parameter is first estimated with a unidimensional 3PL model and then plugged into a MIRT software program for the estimation of the remaining model parameters. Such a two-step estimation procedure, however, has been shown to be inferior to simultaneous estimation of all parameters in the 3PL MIRT model through MCMC estimation (Zhang & Stone, 2004).

As a MIRT model, the 3PL testlet model also faces such issues. Although with MCMC algorithm all the model parameters can be estimated simultaneously, in practice the 3PL testlet model often encounters model convergence issue. According to Wainer, Bradlow, and Wang (2007), estimation difficulty of the 3PL testlet model is mainly due to "weak identifiability, that is, presence of multiple different triplets of for which the ICCs are close" (p. 136), and they caution against using the 3PL testlet model in practice. Due to the estimation difficulty of the 3PL testlet model with MCMC algorithm, it is not uncommon for researchers and practitioners to resort to the 2PL testlet model (e.g., Eckes, 2014).

## 2PL testlet model as a constrained bi-factor model

The 2PL testlet model takes the form

$$p_j(\theta_i) = \frac{1}{1+e^{-a_j\left(\theta_i - b_j - \gamma_{id(j)}\right)}} \tag{1}$$

where $p_j(\theta_i)$ s the probability of person *i* giving a correct response to item *j*, $\theta_i$ s person *i*'s latent ability, $\gamma_{id(j)}$ s person *i*'s latent ability on testlet *d*, and $a_j$ nd $b_j$ re the discrimination and difficulty parameters of item *j*. For $\gamma_{id(j)}$ its variance $\sigma^2_{\gamma_{id(j)}}$ ndicates the magnitude of testlet effect, or LID among items within the same testlet.

The bi-factor model (Gibbons & Hedeker, 1992) is given as

$$p_j(\theta_i) = \frac{1}{1+e^{-(a_{jg}\theta_g - b_j + a_{js}\theta_s)}}, \tag{2}$$

where $a_{jg}$ and $a_{js}$ are the item discrimination parameters on the general and specific factors for item *j*, $b_j$ is the intercept parameter for item *j*, and $\theta_{jg}$ and $\theta_{js}$ are examinee *i*'s latent abilities on the general and specific factors for item *j*.

As illustrated by different researchers (e.g, Li, Bolt, & Fu, 2006; Rijmen, 2010), the bi-factor model in equation 2 can be reduced to the 2PL testlet model in equation 1 by constraining the freely estimated discrimination parameter of the secondary factor ( $a_{js}$ ) to be the product of $\sigma_{\gamma_{id(j)}}$ and $a_j$ Consequently, the 2PL testlet model can be viewed as a constrained bi-factor model that can be estimated with limited-information methods commonly applied to confirmatory factor analysis (CCFA; Wirth & Edwards, 2007) models.

## Three estimation methods

So far, the testlet model has been mainly estimated with MCMC algorithms (e.g., Koziol, 2016; Li, Bolt, & Fu, 2006) or MMLE method (e.g., Jiao, Wang, & He, 2013; Li, Li, & Wang, 2010), both of which are full-information estimation methods. Although widely used in CCFA, limited-information methods can be readily employed for estimation of IRT models due to the well-known mathematical equivalence between CCFA and IRT models that do not include the pseudo-guessing parameter (e.g., Kamata & Bauer, 2008; Takane & de Leeuw, 1987). Studies (e.g., Bolt, 2005; Knol & Berger, 1991) have shown that for MIRT model estimation, limited-information methods such as the unweighted least square (ULS) method (McDonald, 1982) implemented in NOHARM (Fraser, 1988) and the generalized least square (GLS) method implemented in both Mplus and LISREL (Jöreskog & Sörbom, 1993) performed as well as, if not better than, the full-information maximum likelihood method (Bock, Gibbons, & Muraki, 1988) implemented in TEST-

FACT (Wilson, Wood, & Gibbons, 1998). As the 2PL testlet model is also a member of the MIRT family, it is expected that limited-information estimation methods should at least estimate the testlet model parameters as accurately as full-information methods do. WLSMV, the default estimation method for categorical data in Mplus, has been shown to perform better than other limited information methods (e.g., Beauducel & Herzberg, 2006; Flora & Curran, 2004; Yang-Wallentin, Jöreskog, & Luo, 2010). As a robust version of the weighted least square (WLS; Muthen, 1984) method, WLSMV is commonly used for CCFA parameter estimation, and is also increasingly used for estimation of UIRT (Paek, Cui, Gubes, & Yang, 2017) and MIRT models (Finch, 2010).

## Simulation studies on testlet model estimation

We located seven simulation studies related to testlet model estimation through a comprehensive literature search[3]. Among those seven, we excluded three as they focused on either multilevel testlet models (Jiao, Kamata, Wang, & Jin, 2012; Jiao & Zhang, 2015) or higher order testlet models (Huang & Wang, 2013). In the following section, we review the remaining four studies.

Bradlow, Wainer, and Wang (1999) conducted a small-scale simulation to check whether their proposed 2PL testlet model could be accurately estimated with the Gibbs sampler. They simulated a test of 60 items (30 independent items and 30 testlet items) taken by 1000 examinees, and manipulated the number of items per testlet (5 and 10) and the variance of the testlet effects (0.5, 1, and 2) in a 2 by 3 simulation study. They found that the item and ability parameters were accurately recovered. However, they only generated one dataset per simulation condition (likely because the computational burden was prohibitive back then).

In a follow-up study, Wang, Bradlow, and Wainer (2002) extended the 2PL testlet model to a more general Bayesian testlet model that can accommodate mixed format tests containing both dichotomous and polytomous items. They conducted a simulation study with nine conditions to investigate whether the general Bayesian testlet model could be accurately estimated. Number of examinees was fixed at 1000 and number of items was fixed at 30 with 12 independent binary items and 18 testlet items. They manipulated number of items per testlet (3, 6, and 9), number of response categories (2, 5, and 10), and variance of the testlet effects (0, 0.5, and 1). It should be noted that they did not use a fully-crossed design, which would result in 27 simulation conditions, but instead adopted a Latin Square design that reduced the number of simulation conditions to nine. They ran five replications within each condition. They found that their general Bayesian testlet model could be accurately estimated with the Gibbs sampler.

Wang and Wilson (2005) conducted a comprehensive simulation study to investigate whether the Rasch testlet model could be accurately estimated with MMLE implemented

[3] We searched Google Scholar and PsycINFO, as well as all major psychometric and methodological journals, using testlet and estimation as the two key words.

in ConQuest (Wu, Adams, & Wilson, 1998). The four manipulated factors include item type (dichotomous, polytomous, and a mixture of both), number of items within each testlet (5 and 10 for dichotomous items; 3 and 6 for polytomous items), sample size (200 and 500), and variance of the testlet effects (0.25, 0.50, 0.75, and 1). The number of items was also varied with the item type: when the item type was dichotomous, the number of items was 40; when it was polytomous, the number of items was 24; with mixed item format, the number of items was 36 (24 dichotomous and 12 polytomous items). One hundred replications were conducted under each simulation condition. They found that all parameters in the Rasch testlet model could be accurately estimated with MMLE.

In the only study on comparison of estimation methods for testlet models, Jiao, Wang, and He (2013) compared three estimation methods for the one-parameter-logistic (1PL) testlet model. The three methods were MCMC (implemented in WinBUGS; Lunn, Thomas, Best, & Spiegelhalter, 2000), MMLE (implemented in ConQuest), and the sixth-order Laplace approximation (implemented in HLM6; Raudenbush, Bryk, Cheong, & Congdon, 2004). With number of examinees fixed at 1,000 and number of items fixed at 54 (nine items within each of the six testlets), they manipulated the variance of testlet effects (0.25, 0.5625, and 1) and generated 25 datasets under each simulation condition. They found that choice of the estimation method significantly affected parameter recoveries. Specifically, the Laplace method resulted in the smallest bias of testlet variance estimation and the smallest random error of ability parameter estimation, the MMLE method resulted in the smallest bias of ability variance estimation, and the MCMC method resulted in the smallest bias of item difficulty estimation.

To sum up, while there are studies that either focus on one specific estimation method or compare different estimation methods (MCMC vs MMLE) for the 1PL testlet model, none has investigated the performance of limited-information methods for parameter estimation of any testlet model, nor has any study compared the performances of MCMC and MMLE for the 2PL testlet model. Consequently, it remains unknown how the three estimation methods perform regarding model parameter recovery for the 2PL testlet model.

## Methods

### Simulation design

We conducted a Monte Carlo simulation study to compare how well MCMC, MMLE, and WLSMV estimation methods recovered the item parameters for the 2PL testlet model. Number of items was fixed to 30 to mimic a test of medium length. Manipulated factors included sample size (SS; 500, 1000, 2000) and variance of the testlet effects (TV; 0.25, 0.5, 1), which resulted in a fully-crossed design with nine simulation conditions. We set the number of items within each testlet to five to mimic the typical number of items nested within a reading comprehension passage, resulting in six testlets for each dataset.

## Data generation

Similar to Bradlow, Wainer, and Wang (1999), item discrimination and difficulty parameters were generated from a normal distribution $N(1, 0.2)$ and a standard normal distribution $N(0, 1)$, respectively. Table 1 lists the generated item parameters that were used across all simulation conditions. The latent trait parameters were generated from a standard normal distribution $N(0, 1)$, and the person specific testlet effects were generated from a normal distribution with a mean of zero and a variance equal to one of the three possible testlet variance values determined by the specific simulation condition. Note that for the same sample size, the same set of latent ability parameter values was used for data generation. In addition, as all three estimation methods impose a standard normal distribution on the latent ability, the mean and the variance for each of the three sets of generated latent values were exactly zero and one in order for parameter estimates to be on the same metric as the generating parameters without further adjustments. One hundred datasets were generated based on equation 1 for each simulation condition.

**Table 1:**
Item Parameter Values Used for Data Generation

| Item | a | b | Item | a | b |
|---|---|---|---|---|---|
| 1 | 1.17 | -1.55 | 16 | 0.76 | 0.35 |
| 2 | 0.61 | -1.29 | 17 | 0.98 | -1.24 |
| 3 | 0.67 | 1.44 | 18 | 0.70 | 1.30 |
| 4 | 1.16 | 1.86 | 19 | 0.90 | 0.83 |
| 5 | 1.06 | -0.90 | 20 | 0.58 | 0.06 |
| 6 | 0.69 | 0.05 | 21 | 0.66 | -0.41 |
| 7 | 0.81 | -0.88 | 22 | 0.84 | 1.09 |
| 8 | 0.95 | -0.62 | 23 | 0.81 | 0.01 |
| 9 | 0.51 | 1.89 | 24 | 0.77 | -1.06 |
| 10 | 0.88 | 0.09 | 25 | 0.50 | 0.89 |
| 11 | 0.78 | 0.20 | 26 | 0.97 | 0.62 |
| 12 | 0.96 | -0.19 | 27 | 0.62 | -0.17 |
| 13 | 1.21 | 1.89 | 28 | 0.70 | -0.81 |
| 14 | 0.90 | -0.50 | 29 | 0.82 | -0.12 |
| 15 | 0.94 | 0.27 | 30 | 0.77 | -0.43 |

**Estimation procedure**

The Bayes, MLR, and WLSMV estimators in Mplus were used for each estimation method (MCMC, MMLE, and WLSMV, respectively). In Mplus, the default link function for the MLR estimator is a logit link, and we changed it to probit to be consistent with WLSMV and Bayes. When Bayes was specified as the estimator, the following default uninformative priors implemented in Mplus were used: a normal distribution $N(0,5)$ for both the factor loading and item threshold parameters and an inverse gamma distribution $IG(-1,0)$ for the testlet variance parameter. As indicated by Asparouhov and Muthén (2010), $IG(-1,0)$ is equivalent to a uniform distribution $unif(0, \infty)$ and hence uninformative. We specified Mplus to run a minimum of 20,000 iterations for each of four chains so that if the model did not converge after 20,000 iterations, Mplus would keep running the four chains until convergence. Convergence was assessed with a convergence diagnostic index called the potential scale reduction factor (PSRF; Gelman & Rubin, 1992).

As Mplus does not produce IRT-based parameter estimates for the 2PL testlet model, we used the same procedures demonstrated by Luo (2018) to convert the factor loading and item threshold estimates in the Mplus outputs to the corresponding IRT model parameters. Specifically, for the WLSMV estimator, the following two equations (McDonald, 1999) were used to convert the estimated parameters in Mplus output:

$$a_j = \frac{1.702\lambda_{jWLSMV}}{\sqrt{1-\lambda_{jWLSMV}'\phi\lambda_{jWLSMV}}}, \tag{3}$$

$$b_j = \frac{-1.702\tau_{jWLSMV}}{a_j\sqrt{1-\lambda_{jWLSMV}'\phi\lambda_{jWLSMV}}}, \tag{4}$$

where $a_j$ and $b_j$ are the discrimination and difficulty parameters of item $j$, $\lambda_{jWLSMV}$ and $\tau_{jWLSMV}$ are the vector of factor loading parameters and item threshold parameter estimated with the WLSMV estimator, and $\phi$ is the factor covariance matrix. The constant 1.702 is used to convert item parameters from the normal metric to the logistic metric.

For the MLR and Bayes estimators, before equations 3-4 can be used for conversion, the estimates of factor loadings $\lambda_{jBayes/MLR}$ and item threshold $\tau_{jBayes/MLR}$ need to be converted to $\lambda_{jWLSMV}$ and $\tau_{jWLSMV}$ using $R_j^2$, which is the proportion of variance in item $j$ that is accounted for by the latent factor:

$$\lambda_{jWLSMV} = \lambda_{jBayes/MLR}\sqrt{1-R_j^2}, \tag{5}$$

$$\tau_{jWLSMV} = \tau_{jBayes/MLR}\sqrt{1-R_j^2}. \tag{6}$$

### Outcome variables

To evaluate the ability of each estimation method to accurately recover the item and testlet variance parameters, we examined bias, standard error (SE), and root mean square error (RMSE). Bias, SE, and RMSE are indicative of systematic error, random error, and total error of estimation, respectively. They are defined as

$$Bias(\hat{\pi}) = \frac{\Sigma_1^R(\hat{\pi}_r - \pi)}{R}, \tag{7}$$

$$SE(\hat{\pi}) = \sqrt{\frac{\Sigma_1^R(\hat{\pi}_r - \bar{\hat{\pi}})^2}{R}} \tag{8}$$

and

$$RMSE(\hat{\pi}) = \sqrt{\frac{\Sigma_1^R(\hat{\pi}_r - \pi)^2}{R}} \tag{9}$$

where $\pi$ s the true model parameter, $\hat{\pi}_r$ s the estimated model parameter for the $r$th replication, $\bar{\hat{\pi}}$ s the mean of model parameter estimates across replications, and $R$ is the number of replications.

We further examined the mean relative bias (RB) of model parameter recovery, the value of which can be used to gauge the magnitude of estimation bias. RB is also called percentage bias and can be computed as

$$RB(\hat{\pi}) = \frac{\Sigma_1^R \frac{\hat{\pi}_r - \pi}{\pi}}{R} \tag{10}$$

where all the terms remain the same as in equations 7-9. Previous studies (e.g., Flora & Curran, 2004; Kaplan, 1989) treat parameter estimates with RB values no greater than five as trivially biased, between five and ten as moderately biased, and greater than 10 as substantially biased.

## Results

### Model convergence

The Bayes estimator had no convergence issues regardless of sample size and testlet variance magnitude. Thus, this section focuses on the WLSMV and MLR estimators. For the WLSMV estimator, the occurrence of Heywood cases (negative residual variance estimates) is indicative of model non-convergence; for the MLR estimator, Mplus generates a warning message stating that the model did not converge. It should be noted that when Heywood cases occur with WLSMV, Mplus still produces parameter estimates but

they cannot be trusted; when the model does not converge with MLR, Mplus does not produce any estimates at all. For all the analyses regarding model parameter recovery in the following sections, only those datasets without estimation convergence issues were used for the computation of bias, SE, RMSE and RB.

When sample size was small (SS = 500) and variance of the testlet effects small (TV = 0.25), in 26 replications the WLSMV estimator encountered Heywood cases, and in 10 replications the MLR estimator failed to converge. With the same sample size but medium testlet variance (TV = 0.50), Heywood cases occurred in 10 replications under the WLSMV estimator, while model non-convergence did not occur at all for the MLR estimator. When sample size was medium (SS = 1,000) and variance of the testlet effects small (TV = 0.25), Heywood cases occurred in 10 replications under the WLSMV estimator, and the MLR estimator had no model non-convergence issues. When sample size was large (SS = 2000), neither WLSMV nor MLR encountered model non-convergence regardless of magnitude of the testlet variance. We conclude that for the 2PL testlet model, both WLSMV and MLR tend to have model convergence issues with small sample sizes and small variances of the testlet effects, but the increase of either sample size or testlet variance reduces the occurrence of model non-convergence for both estimators.

## Computation time

To run our simulations, we used a desktop computer equipped with an eight-core 2.40 GHz Xeon processor. As the Bayes estimator in Mplus implements the often time-consuming Gibbs sampler and the 2PL testlet model is fairly complex, we had expected that the computation time with the Bayes estimator would be much longer than with both

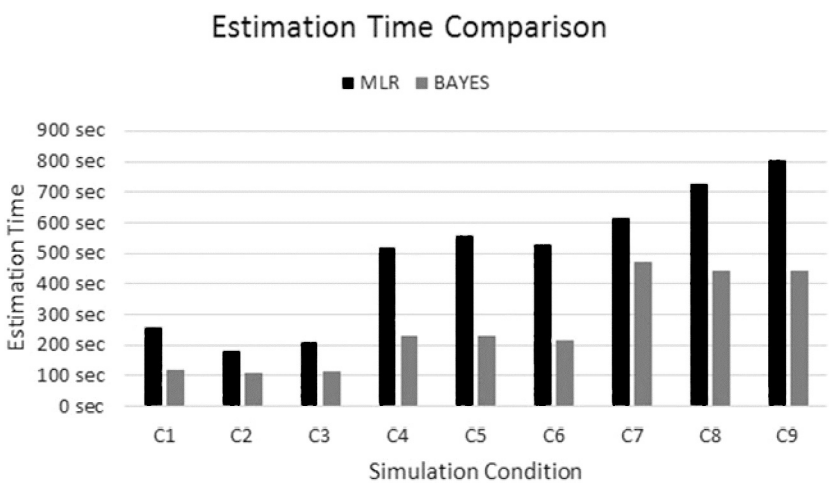


**Figure 1**:
Estimation time comparison between MLR and BAYES estimators in Mplus

the MLR and WLSMV estimators. However, as indicated by the average computation time comparison between MLR and Bayes in Figure 1, MLR took considerably more time than Bayes did across all nine simulation conditions. WLSMV is not included here, as its estimation time ranged from one to two seconds.

## Testlet variance recovery

Table 2 lists the descriptive statistics of estimation biases for the testlet variance parameter. To evaluate the effect of different manipulated factors on estimation biases, we utilized a three-way analysis of variance (ANOVA) model with bias as the dependent variable and estimation method, sample size, and testlet variance magnitude as the three independent variables. The results indicated that the main effect of estimation method was not statistically significant, $F$ (2,135) = 2.576, $p$ = .080. In other words, the three estimation methods resulted in comparable testlet variance estimates.

The main effect of sample size was statistically significant, $F$ (2,135) = 9.470, $p$ = .000, $f$ = .375. Tukey *post hoc* tests revealed that estimation biases of the testlet variance with SS = 500 were significantly greater than those with SS = 1,000 ($p$ = 0.001) and SS = 2,000 ($p$ = 0.001), while differences in estimation biases between SS = 1,000 and SS = 2,000 were insignificant. The magnitude of testlet variance was also statistically significant, $F$ (2,135) = 8.172, $p$ =.000, $f$ = .348. Tukey *post hoc* tests revealed that estimation biases of testlet variance when testlet variances were generated to be large were significantly greater than those with medium ($p$ = 0.001) and small testlet variance ($p$ = 0.003), while there were no significant differences between small and medium testlet variances.

The interaction between sample size and magnitude of testlet variance was also statistically significant, $F$ (4,135) = 14.600, $p$ = .000, $f$ = .658. This significant interaction effect indicates how estimation bias of the testlet variance changes with the testlet variance

**Table 2:**
Bias in Testlet Variance Estimation

| SS | TV | WLSMV | | MLR | | BAYES | |
|---|---|---|---|---|---|---|---|
| | | Mean | SD | Mean | SD | Mean | SD |
| 500 | L | .0071 | .0042 | .0092 | .0042 | .0104 | .0039 |
| | M | -.0021 | .0025 | -.0014 | .0022 | .0005 | .0021 |
| | S | .0008 | .0064 | -.0008 | .0068 | .0026 | .0057 |
| 1000 | L | -.0007 | .0047 | .0006 | .0047 | .0011 | .0047 |
| | M | -.0004 | .0016 | .0002 | .0015 | .0011 | .0014 |
| | S | -.0004 | .0031 | -.0009 | .0033 | .0011 | .0030 |
| 2000 | L | -.0012 | .0042 | -.0006 | .0039 | -.0007 | .0038 |
| | M | .0009 | .0030 | .0013 | .0029 | .0014 | .0029 |
| | S | .0002 | .0021 | .0004 | .0022 | .0007 | .0021 |

magnitude depends on the sample size. This pattern can be observed in the upper left panel of Figure 2. The marginal estimation bias (averaged across three estimation methods) with SS = 500 drops considerably when the testlet variance magnitude changes from large to medium, while the marginal estimation bias increases slightly from large to medium testlet variance with SS = 2,000. However, the marginal estimation bias remains approximately the same regardless of the magnitude of the testlet variance with SS = 1,000.

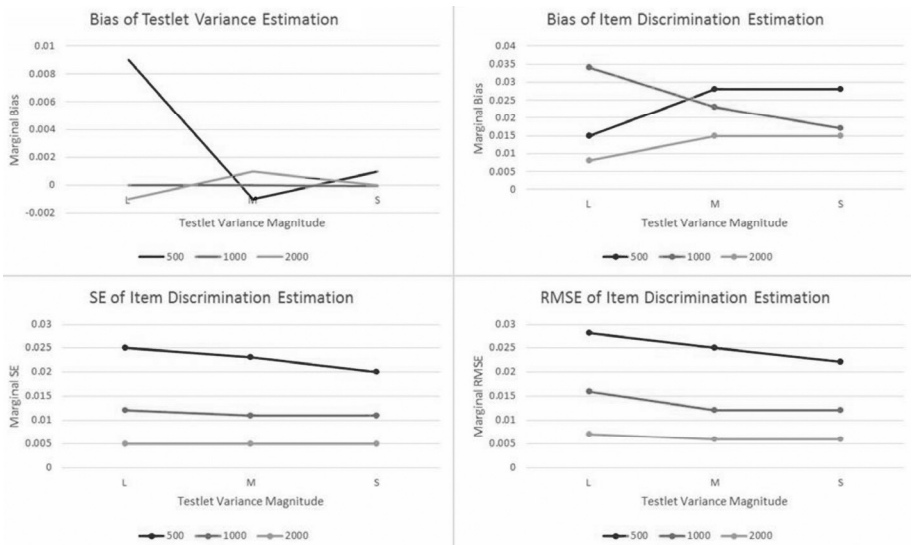


**Figure 2:**
Significant interaction effects between sample size and testlet variance magnitude

## Item difficulty parameter recovery

Estimation biases, SEs, and RMSEs of the item difficulty parameter were summarized in Table 3. Using a similar three-way ANOVA model, we again found that estimation method was not a significant predictor of bias, $F$ (2,783) = .168, $p$ = .845, indicating that the three estimation methods produced item difficulty estimates of comparable biases. The main effect of magnitude of testlet variance was also insignificant, $F$ (2,783) = .227, $p$ = .797. The main effect of sample size was statistically significant, $F$ (2,783) = 3.711, $p$ = .025, $f$ = .095. Tukey *post hoc* tests revealed that estimation biases of item difficulty with SS = 500 was significantly greater than those with SS = 1,000 ($p$ = .027) but not statistically different from those with SS = 2,000 ($p$ = .101). There were no significant differences in estimation biases between SS = 1,000 and SS = 2,000 ($p$ = .858).

**Table 3:**
Recovery of Item Difficulty Parameter

| Estimator | SS | TV | Bias | | SE | | RMSE | |
|---|---|---|---|---|---|---|---|---|
| | | | Mean | SD | Mean | SD | Mean | SD |
| Bayes | 500 | L | .0213 | .0847 | .0490 | .0536 | .0564 | .0611 |
| | | M | .0061 | .0374 | .0421 | .0438 | .0435 | .0447 |
| | | S | .0107 | .0464 | .0455 | .0453 | .0477 | .0477 |
| | 1000 | L | -.0057 | .0571 | .0181 | .0140 | .0212 | .0172 |
| | | M | .0042 | .0383 | .0204 | .0270 | .0218 | .0299 |
| | | S | .0074 | .0499 | .0254 | .0404 | .0278 | .0500 |
| | 2000 | L | .0008 | .0555 | .0090 | .0085 | .0120 | .0138 |
| | | M | .0037 | .0230 | .0095 | .0103 | .0100 | .0108 |
| | | S | .0049 | .0282 | .0091 | .0084 | .0099 | .0096 |
| MLR | 500 | L | .0215 | .0896 | .0489 | .0536 | .0572 | .0617 |
| | | M | .0063 | .0423 | .0421 | .0436 | .0438 | .0451 |
| | | S | .0118 | .0514 | .0453 | .0458 | .0480 | .0490 |
| | 1000 | L | -.0069 | .0593 | .0180 | .0140 | .0215 | .0175 |
| | | M | .0033 | .0378 | .0203 | .0269 | .0217 | .0298 |
| | | S | .0060 | .0504 | .0251 | .0396 | .0276 | .0488 |
| | 2000 | L | .0010 | .0578 | .0090 | .0086 | .0123 | .0143 |
| | | M | .0038 | .0253 | .0095 | .0103 | .0102 | .0110 |
| | | S | .0051 | .0302 | .0091 | .0085 | .0100 | .0098 |
| WLSMV | 500 | L | .0216 | .0832 | .0525 | .0628 | .0596 | .0691 |
| | | M | .0009 | .0345 | .0419 | .0412 | .0430 | .0418 |
| | | S | .0059 | .0428 | .0427 | .0441 | .0445 | .0464 |
| | 1000 | L | -.0085 | .0562 | .0184 | .0147 | .0215 | .0174 |
| | | M | .0021 | .0355 | .0202 | .0271 | .0214 | .0292 |
| | | S | .0041 | .0454 | .0245 | .0382 | .0265 | .0462 |
| | 2000 | L | -.0004 | .0538 | .0096 | .0101 | .0124 | .0145 |
| | | M | .0034 | .0232 | .0098 | .0110 | .0103 | .0115 |
| | | S | .0040 | .0263 | .0091 | .0082 | .0097 | .0092 |

*Note.* TV stands for testlet variance, SS for sample size; L stands for Large, M for Medium, S for small.

The results of a three-way ANOVA model with SE as the dependent variable indicated that the three estimation methods did not differ statistically, $F$ (2,783) = .001, $p$ = .999, which means that item difficulty estimates had similar SEs regardless of the estimation method. The main effect of magnitude of testlet variance was also insignificant, $F$

(2,783) = .348, $p$ = .706. The main effect of sample size was statistically significant, $F$ (2,783) = 84.200, $p$ = .000, $f$ =.464. Tukey *post hoc* tests revealed that SEs of item difficulty estimation with SS = 500 were significantly greater than those with SS=1,000 ($p$ = .000) and SS = 2,000 ($p$ = .000); SEs of item difficulty with SS = 1,000 were also significantly greater than those with SS = 2,000 ($p$ = .000).

Finally, the results of a three-way ANOVA with RMSE as the dependent variable indicated that the three estimation methods were not statistically significantly different, $F$ (2,783) = .007, $p$ = .993. In other words, estimation method did not affect RMSEs of item difficulty estimates. The magnitude of testlet variance was also insignificant, $F$ (2,783) = 1.429, $p$ = .240. The main effect of sample size was statistically significant, $F$ (2,783) = 76.501, $p$ = .000, $f$ = .441. Tukey *post hoc* tests revealed that RMSEs of item difficulty with SS = 500 were significantly greater than those with SS = 1,000 ($p$ = .000) and SS = 2,000 ($p$ = .000); RMSEs of item difficulty with SS = 1,000 were also significantly greater than those with SS =2,000 ($p$ = .000).

**Item discrimination parameter recovery**

Estimation biases, SEs, and RMSES of the item discrimination parameter are summarized in Table 4. The results of a three-way ANOVA with bias as the dependent variable indicated that estimation method as a main effect was not statistically significant, $F$ (2,783) = 1.859, $p$ = .156. In other words, the three estimation methods produced item discrimination estimates of comparable biases. The main effect of magnitude of testlet variance was also insignificant, $F$ (2,783) = .522, $p$ = .593. The main effect of sample size was statistically significant, $F$ (2,783) = 7.233, $p$ = .001, $f$ = .135. Tukey *post hoc* tests revealed that estimation biases of item discrimination with SS = 2,000 were significantly smaller than those with SS = 1,000 ($p$ = .002) and SS = 500 ($p$ = .005), and there were no significant differences in estimation biases between SS = 1,000 and SS = 500 ($p$ = .962). The interaction between sample size and magnitude of testlet variance was also statistically significant, $F$ (4, 783) = 4.087, $p$ = .003, $f$ = .143. As can be seen in the upper right panel of Figure 2, this significant interaction effect indicates how testlet variance magnitude affects estimation bias of the item discrimination parameter depends on sample size. However, as indicated by Cohen's $f$, this interaction effect has a considerably smaller effect size than its counterpart in the ANOVA model with testlet variance estimation bias as the dependent variable.

The results of a three-way ANOVA with SE as the dependent variable indicated that the main effect of estimation method was not statistically significant, $F$ (2,783) = 1.388, $p$ = .250, which means that item discrimination estimates had similar SEs regardless of estimation method. The main effect of sample size was statistically significant, $F$ (2,783) = 492.283, $p$ = .000, $f$ = 1.121. Tukey *post hoc* tests revealed that SEs of item discrimination estimation with SS = 500 were significantly greater than those with SS = 1,000 ($p$ = .000) and SS = 2,000 ($p$ = .000); SEs of item discrimination estimation with SS = 1,000 were also significantly greater than those with SS = 2,000 ($p$ = .000). The main effect of

**Table 4:**
Parameter Recovery for Item Discrimination

| Estimator | SS | TV | Bias | | SE | | RMSE | |
|---|---|---|---|---|---|---|---|---|
| | | | Mean | SD | Mean | SD | Mean | SD |
| Bayes | 500 | L | .0171 | .0575 | .0249 | .0143 | .0284 | .0161 |
| | | M | .0321 | .0383 | .0236 | .0102 | .0261 | .0106 |
| | | S | .0316 | .0338 | .0211 | .0076 | .0233 | .0083 |
| | 1000 | L | .0336 | .0564 | .0113 | .0053 | .0155 | .0073 |
| | | M | .0233 | .0278 | .0110 | .0048 | .0123 | .0055 |
| | | S | .0174 | .0285 | .0108 | .0045 | .0119 | .0050 |
| | 2000 | L | .0086 | .0418 | .0051 | .0018 | .0069 | .0045 |
| | | M | .0161 | .0300 | .0053 | .0019 | .0064 | .0028 |
| | | S | .0163 | .0319 | .0049 | .0019 | .0061 | .0034 |
| MLR | 500 | L | .0080 | .0567 | .0229 | .0121 | .0261 | .0142 |
| | | M | .0221 | .0405 | .0219 | .0082 | .0239 | .0091 |
| | | S | .0230 | .0362 | .0196 | .0063 | .0214 | .0075 |
| | 1000 | L | .0303 | .0577 | .0110 | .0050 | .0151 | .0070 |
| | | M | .0198 | .0282 | .0107 | .0045 | .0119 | .0053 |
| | | S | .0139 | .0298 | .0105 | .0043 | .0116 | .0051 |
| | 2000 | L | .0053 | .0421 | .0051 | .0017 | .0068 | .0046 |
| | | M | .0131 | .0301 | .0052 | .0019 | .0063 | .0029 |
| | | S | .0134 | .0321 | .0048 | .0019 | .0060 | .0035 |
| WLSMV | 500 | L | .0184 | .0572 | .0264 | .0132 | .0299 | .0148 |
| | | M | .0308 | .0389 | .0230 | .0081 | .0255 | .0086 |
| | | S | .0300 | .0307 | .0198 | .0077 | .0216 | .0082 |
| | 1000 | L | .0375 | .0584 | .0122 | .0056 | .0169 | .0078 |
| | | M | .0259 | .0269 | .0112 | .0047 | .0126 | .0054 |
| | | S | .0195 | .0279 | .0109 | .0043 | .0120 | .0045 |
| | 2000 | L | .0092 | .0410 | .0057 | .0020 | .0074 | .0042 |
| | | M | .0164 | .0302 | .0056 | .0020 | .0067 | .0028 |
| | | S | .0167 | .0304 | .0049 | .0019 | .0061 | .0031 |

*Note.* TV stands for testlet variance, SS for sample size; L stands for Large, M for Medium, S for small.

magnitude of testlet variance was also statistically significant, $F$ (2,783) = 5.868, $p$ = .003, $f$ = 0.123. Tukey *post hoc* tests revealed that SEs of item discrimination with large testlet variance were significantly greater than those with small testlet variance ($p$ = .002) but not those with medium testlet variance ($p$ = .349), and that there were no statistically

significant differences between medium and small testlet variance ($p$ = .108). The interaction between sample size and magnitude of testlet variance was also statistically significant, $F$ (4, 783) = 2.806, $p$ = .025, $f$ = .119. As seen in the bottom left panel of Figure 2, this significant interaction effect shows how testlet variance magnitude affects estimation SE of the item discrimination parameter depends on sample size, but the effect size is fairly small.

The results of a three-way ANOVA with RMSE as the dependent variable indicated that the main effect of estimation method was not statistically significant, $F$ (2,783) = 1.481, $p$ = .228. In other words, estimation method did not affect RMSEs of item discrimination estimates. The main effect of sample size was statistically significant, $F$ (2,783) = 408.849, $p$ = .000, $f$ = 1.022. Tukey *post hoc* tests revealed that RMSEs of item discrimination estimation with SS = 500 were significantly greater than those with SS = 1,000 ($p$ = .000) and SS = 2,000 ($p$ = .000); RMSEs of item discrimination estimation with SS = 1,000 were also significantly greater than those with SS =2,000 ($p$ = .000). The main effect of magnitude of testlet variance was also statistically significant, $F$ (2,783) = 16.085, $p$ = .000, $f$ = .201. Tukey *post hoc* tests revealed that RMSEs of item discrimination with large testlet variance were significantly greater than those with both small ($p$ = .000) and medium testlet variance ($p$ = .001), and there were no statistically significant differences between medium and small testlet variance ($p$ = .116). The interaction between sample size and magnitude of testlet variance was also statistically significant, $F$ (4, 783) = 2.958, $p$ = .019, $f$ = .123. As seen in the bottom right panel of Figure 2, this significant interaction effect with a small effect size shows that the relation between magnitude of testlet variance and estimation RMSEs of the item discrimination parameter is slightly moderated by sample size.

**Relative bias of model parameter estimates**

The mean RB values for model parameter estimates in each of the simulation conditions are listed in Table 5. Specifically, the numbers under the column *a* are the mean RB values for item discrimination parameter estimates, those under the column *b* are for item difficulty parameter estimates, and those under the column $\sigma^2$ are for testlet variance estimates. As can be seen, most of the RB values for the item discrimination estimates and the testlet variance estimates are smaller than five, suggesting that they have trivial biases across different simulation conditions regardless of estimation method. The mean RB values for the item difficulty parameter estimates are greater than five in some conditions. After closely examining the RB values for each of the thirty item difficulty parameter estimates in each simulation condition, we conclude that their larger mean RB values are merely a function of the small generating values of two items (the generating *b* values for item 6 and 23 are 0.05 and 0.01, respectively) and the way RB is computed. That is, as the generating values are used as the denominator in the computation of RB, RB values tend to become large with small generating values even when bias is small. However, such large values do not indicate that the item difficulty estimates have moderate biases.

**Table 5:**
RB Values for Model Parameter Estimates

| SS | TV | Bayes | | | MLR | | | WLSMV | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | ***a*** | ***b*** | $\sigma^2$ | ***a*** | ***b*** | $\sigma^2$ | ***a*** | ***b*** | $\sigma^2$ |
| 500 | L | 2.29 | 2.70 | 1.04 | 1.28 | 3.75 | .92 | 2.47 | 2.59 | .71 |
| | M | 4.41 | -4.70 | .09 | 3.28 | -3.68 | -.25 | 4.33 | -3.85 | -.37 |
| | S | 4.20 | 2.72 | 1.02 | 3.23 | 3.52 | -.32 | 3.98 | 2.85 | .32 |
| 1000 | L | 4.55 | -4.58 | .11 | 4.19 | -4.72 | .06 | 5.01 | -5.19 | -.07 |
| | M | 3.14 | 2.58 | .20 | 2.75 | 2.46 | .03 | 3.47 | 1.74 | -.07 |
| | S | 2.50 | 5.99 | .45 | 2.10 | 5.83 | -.37 | 2.75 | 5.16 | -.16 |
| 2000 | L | 1.44 | 4.80 | -.07 | 1.03 | 5.30 | -.06 | 1.50 | 5.15 | -.12 |
| | M | 2.40 | -2.73 | .25 | 2.03 | -2.39 | .23 | 2.43 | -2.68 | .15 |
| | S | 2.41 | -.34 | .29 | 2.07 | -.07 | .15 | 2.44 | -.26 | .07 |

*Note.* TV stands for testlet variance, SS for sample size; L stands for Large, M for Medium, S for small.

Figure 3 provides a visual summary of the mean RB value comparison averaged across nine simulation conditions for the three estimation methods. As can be seen, all three estimation methods have mean RB values well below the threshold of five, suggesting that on average, the three estimation methods produce model parameter estimates of trivial biases.

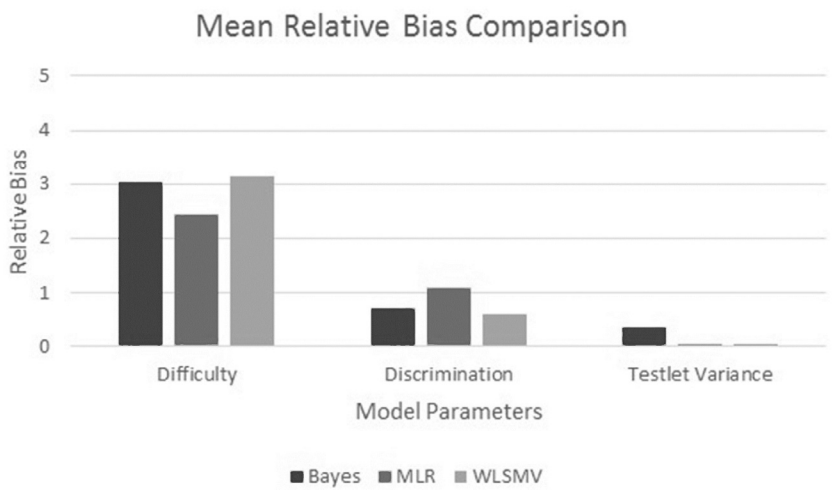


**Figure 3:**
Mean RB values across nine simulation conditions

## A real data example

In this section we use a real dataset to demonstrate the estimation of the 2PL testlet model with the three different methods. The data come from student responses to a test form of the Verbal Section of the General Aptitude Test (GAT-V), a high-stakes test used for college admission purpose in Saudi Arabia. Each GAT-V test has 52 multiple-choice items, and 19 or 20 of those are reading comprehension items nested within four passages. In the current GAT-V test form, there are 19 reading comprehension items, with the third passage having four items, and the other 15 items evenly distributed across the three remaining passages. We drew a random sample of 2,000 students who took the current test form and extracted their responses to the 19 reading comprehension items. The subsequent analyses were based on this 2,000 by 19 response matrix of zeros and ones, where one indicates a correct response and zero an incorrect response.

We fit the 2PL testlet model to this dataset using the Bayes, MLR, and WLSMV estimators (for Mplus syntax on how to estimate the 2PL testlet model with different estimation methods, see, Luo, 2018). Note that for the BAYES estimator that implements the MCMC algorithm, we specified Mplus to run four parallel chains with each one containing a minimum of 20,000 iterations. The iteration history shows that the largest PSRF value dropped below 1.1 after 9,000 iterations and at 20,000 iterations, the largest PSRF value was just 1.027, suggesting that model convergence was reached. In addition, the Bayesian posterior predictive p (PPP) value is 0.173, indicating an excellent data fit of the 2PL testlet model. In terms of computation time, WLSMV took one second, Bayes 313 seconds, and MLR 322 seconds on the same desktop computer equipped with an eight-core Xeon 2.4 GHz processor.

Table 6 lists the testlet variance estimates for the four passages based on the three estimation methods. All methods indicate that the first two passages have small testlet variances, the third passage a negligible testlet variance, and the fourth passage a moderate testlet variance.

Figure 4 provides a visual presentation of comparisons of the item parameters estimates (converted via equations 3-6) based on the three estimation methods. Each of the six graphs in Figure 4 represents a compression between two sets of item parameter estimates based on two estimation methods, and the red dotted regression line $y = x$ in each graph is superimposed to show how different the two sets of estimates are.

**Table 6:**
Testlet Variance Estimates for the Real Data

| Estimator | Passage 1 | Passage 2 | Passage 3 | Passage 4 |
|---|---|---|---|---|
| WLSMV | 0.089 | 0.166 | 0.033 | 0.635 |
| MLR | 0.095 | 0.176 | 0.036 | 0.518 |
| BAYES | 0.104 | 0.181 | 0.058 | 0.557 |

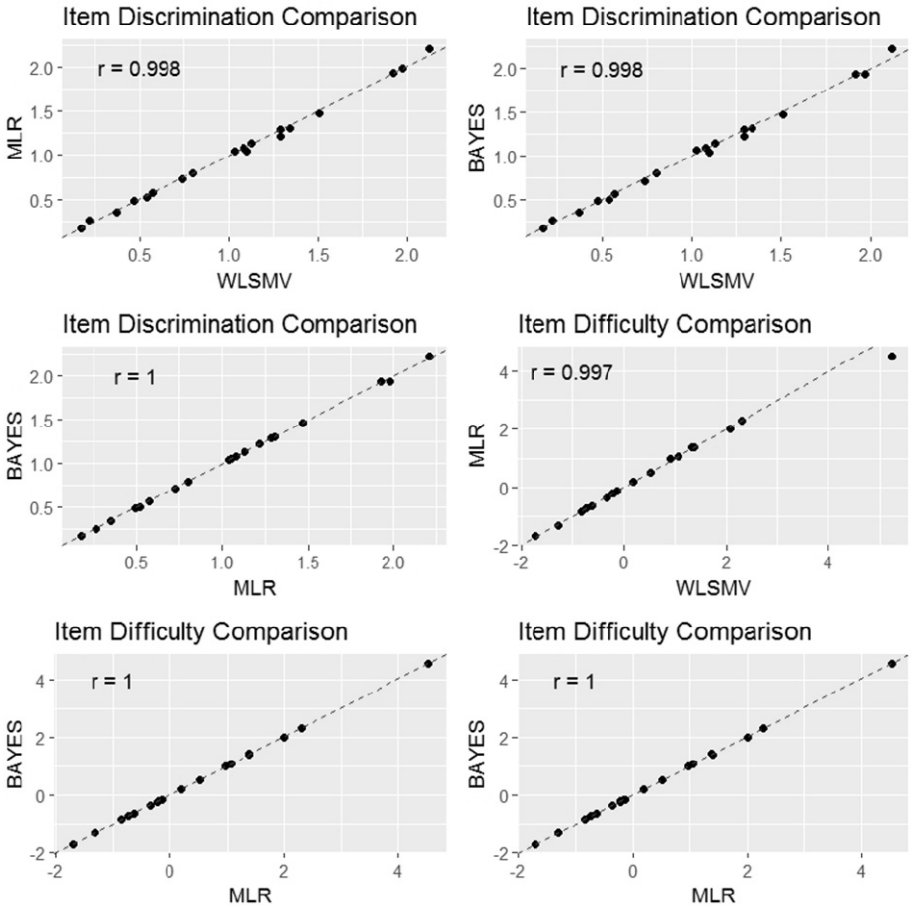


**Figure 4:**
Item parameter estimates comparison

As can be seen, the three sets of converted item parameters are highly similar, with the lowest correlation coefficient being 0.997. In addition, most of the points land perfectly on the regression line $y = x$, suggesting that item parameter estimates based on the three estimation methods are almost identical. We conclude that regardless of the choice of estimation method, the testlet variance estimates and the item difficulty and discrimination parameter estimates for the current dataset are virtually the same, corroborating the findings of the simulation study that item parameter estimates based on the three estimation methods are comparable.

## Conclusions and discussions

The 2PL testlet model is a popular member of the testlet model family commonly used by researchers and practitioners to address the issue of LID. As a constrained bifactor model, the 2PL testlet model can be estimated with both full information and limited information methods. However, no studies exist in the psychometric literature that systematically compare the performances of different estimation methods. Intended to fill the gap in literature, the current study investigated the performances of MCMC, MMLE, and WLSMV as estimation methods for the 2PL testlet model via a simulation study.

Utilizing the large number of estimation methods available in Mplus, we implemented the three estimation methods within Mplus and eliminated the potential confounding effect caused by use of different software programs. The most important finding of our simulation study was that estimation method did not significantly affect bias, SE, or RMSE of testlet variance estimation, or those of item discrimination and difficulty parameter estimation. In addition, parameter estimates were only trivially biased regardless of estimation method. In other words, all three estimation methods performed similarly and accurately when estimating each parameter.

The results also shed light on how sample size and testlet variance magnitude affect the item parameter recovery of the 2PL testlet mode. Specifically, sample size had significant effect on bias of testlet variance parameter estimation. Increasing the sample size from 500 to 1,000 resulted in significantly smaller biases, but from 1,000 to 2,000 did not reduce the bias significantly. Regarding the estimation of item difficulty and discrimination parameters, sample size significantly affected their estimation bias, SE, and RMSE. Increasing the sample size from both 500 to 1,000 and from 1,000 to 2,000 generally seemed to help improve the estimation quality of both difficulty and discrimination. Magnitude of testlet variance had a significant effect on bias of the testlet variance parameter estimation, and SE and RMSE of the item discrimination parameter estimation. For estimation of the testlet variance parameter, decreasing its magnitude from large to medium did not reduce the bias significantly, but decreasing its magnitude from medium to small did lead to significantly smaller biases. For SE of the item discrimination parameter estimation, large testlet variance resulted in significantly smaller SEs than small testlet variance; for RMSE of the item discrimination parameter estimation, large testlet variance led to significantly smaller RMSEs than both small and medium testlet variance. In addition, significant interaction effects between sample size and testlet variance magnitude were found regarding the estimation of testlet variance and item discrimination parameters, indicating that the effects of testlet variance magnitude on recovery of those parameters were moderated by sample size.

The similar performances regarding parameter recovery notwithstanding, the three estimation methods did differ in terms of model convergence rate. MCMC had no issues with model convergence regardless of the sample size and magnitude of testlet variance. When the sample size and the magnitude of testlet variance were small (SS = 500, TV = 0.25), both MMLE and WLSMV encountered model non-convergence issues, and the latter was more likely to do so. With the increase of either the sample size or the magni-

tude of testlet variances, the number of model non-convergence occurrences decreased for WLSMV and disappeared for MMLE.

If model convergence is not an issue, WLSMV seems a promising estimation method considering its comparable performances with MMLE and MCMC and its quick computation. As mentioned previously, for the simulated datasets under the current simulation conditions, it took up to two seconds for WLSMV to estimate the 2PL testlet model. It is worth noting that the computation time of WLSMV is known to be sensitive to the number of items other than the number of observations, and it is expected that for a longer test (with 60 items or more), the computation time will be considerably longer to the extent that the speed advantage of WLSMV may totally disappear. However, for tests of medium length, WLSMV is the fastest among the three. In addition, one potential benefit of using limited-information estimation methods for IRT estimation, as pointed out by McDonald and Mok (1995), is the existence of an abundance of fit statistics such as the comparative fit index (CFI) and the weighted root mean square residual (WRMR) that can help with model fit assessment.

If model convergence is a concern, the MCMC method seems to be an ideal candidate. With the Bayes estimator that implements the MCMC method, Mplus automatically computes the Bayesian PPP value, which can be used for model check purposes; in contrast, the MMLE method provides information criterion based indices such as AIC and BIC, which can only be used for selection of the optimal model among a group of candidates but not for assessment of model fit. However, one potential drawback of the Bayes estimator in Mplus is that unlike the WLSMV and MLR estimators, the latent ability estimate (or factor score) cannot be requested directly. One can circumvent this limitation by requesting Mplus to draw multiple plausible values from the estimated posterior distribution and obtain the corresponding point estimates by computing the posterior mean, which is analogous to Expected A Priori (EAP) estimates obtained with MMLE. It has been shown that twenty plausible values can lead to latent ability estimates as accurate as those obtained via MMLE/EAP (Luo & Dimitrov, 2018).

One limitation of the present study is that we only considered a test of medium length. It may be of interest to increase the number of items to sixty or more to mimic a longer test. We also fixed the number of items within a testlet to five to represent the number of items within reading comprehension passages commonly seen in practice. Future simulation studies should manipulate these two factors to see whether the findings reported in this study are generalizable. In addition, the only limited information estimation method included was WLSMV; there are other promising methods such as the ULS method implemented in NOHARM. One potential direction for future research is to compare the performances of WLSMV and ULS in NOHARM as estimation methods for testlet models.

To conclude, the findings of the current study have important implications for users of testlet models. First, the simulation results showed that the three estimation methods are all viable estimation methods for the 2PL testlet model, and comparable results can be gained regardless of the choice of estimation method. Second, the current study demonstrated that Mplus, with its provision of various estimation methods and easy-to-learn

syntax, is a competitive candidate when it comes to the choice of software for the 2PL testlet model estimation. For those who prefer to use MCMC for the estimation but are concerned with the slow computation of MCMC in Bayesian software such as WinBUGS and its potential non-convergence issue, the Gibbs sampler implemented in Mplus turns out to be quicker than MMLE. While Stan can be another appealing option for estimation of the 2PL testlet model due to its efficiency in estimating complex IRT models (e.g., Luo & Jiao, 2018), computation of the Bayesian PPP value in Stan is not automatic and requires a considerable amount of programing, a potential hurdle for applied researchers who may not be familiar with the Stan syntax enough to program such posterior predictive check procedures.

## References

Ackerman, T. (1987). *The robustness of LOGIST and BILOG IRT estimation programs to violations of local independence*. ACT Research Report Series, 87–14. Iowa City, IA: American College Testing.

Asparouhov, T., & Muthén, B. (2010). *Bayesian analysis of latent variable models using Mplus*. Technical Report.

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling, 13*(2), 186-203.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443-459.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, *12*(3), 261-280.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*(2), 153-168.

Braeken, J., Tuerlinckx, F., & De Boeck, P. (2007). Copula functions for residual dependency. *Psychometrika*, *72*(3), 393-411.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., ... & Riddell, A. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software.*

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

Finch, H. (2010). Item parameter estimation for the MIRT model: Bias and precision of confirmatory factor analysis-based models. *Applied Psychological Measurement*, *34*(1), 10-26.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*(4), 466-491.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 457-472.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*(3), 423-436.

Hoffman, M. D., & Gelman, A. (2014). The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(1), 1593-1623.

Hoskens, M., & De Boeck, P. (1997). A parametric model for local dependence among test items. *Psychological methods*, *2*(3), 261.

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.

Ip, E. H. (2002). Locally dependent latent trait model and the Dutch identity revisited. *Psychometrika*, *67*(3), 367-386.

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.

Jiao, H., & Zhang, Y. (2015). Polytomous multilevel testlet models for testlet-based assessments with complex sampling designs. *British Journal of Mathematical and Statistical Psychology, 68*(1), 65-83.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: User's reference guide*. Chicago: Scientific Software International.

Kaplan, D. (1989). A study of the sampling variability and z-values of parameter estimates from misspecified structural equation models. *Multivariate Behavioral Research*, *24*(1), 41-57.

Knol, D. L., & Berger, M. P. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research, 26(*3), 457-477.

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.

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 RR-10-21). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS – a Bayesian modeling framework: Concepts, structure, and extensibility. *Statistics and Computing, 10*, 325– 337.

Luo, Y. (2018). A short note on estimating the testlet model using different estimators in Mplus. *Educational and Psychological Measurement*, *78*(3), 517-529.

Luo, Y., & Dimitrov, D. M. (2018). A short note on obtaining point estimates of the IRT ability parameter with MCMC estimation in Mplus: how many plausible values are needed? *Educational and Psychological Measurement*.

Luo, Y., & Jiao, H. (2018). Using the Stan program for Bayesian item response theory. *Educational and Psychological Measurement*, *78*(3), 384-408.

McDonald, R. P. (1999). *Test theory: A unified approach.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

McDonald, R. P., & Mok, M. M. C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research*, *30*(1), 23-40.

McLeod, L. D, Swygert, K. A., & Thissen, D. (2001). Factor analysis for items scored in two categories. In D. Thissen and H. Wainer (Eds.), *Test scoring* (pp. 189-216). Mahwah, NJ: Lawrence Erlbaum.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika, 43*(4), 551–560.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49*(1), 115–132.

Muthén, B., du Toit, S. H. C., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*. Unpublished manuscript.

Muthén, L., & Muthén, B. (1998-2012). *Mplus user's guide (Seventh Edition)*. Los Angeles, Ca: Muthén & Muthén.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, *2*, 113-162.

Paek, I., Cui, M., Öztürk Gübeş, N., & Yang, Y. (2017). Estimation of an IRT model by Mplus for dichotomously scored responses under different estimation methods. *Educational and Psychological Measurement*, 0013164417715738.

Patz, R. J., & Junker, B. W. (1999a). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*(2), 146-178.

Patz, R. J., & Junker, B. W. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral statistics*, *24*(4), 342-366.

Raudenbush, S. W., Yang, M. L., & Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics, 9*(1), 141-157.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. (2004). *HLM6: Hierarchical linear and nonlinear modeling [Computer program]*. Chicago, IL: Scientific Software International.

Reckase, M. (2009). *Multidimensional item response theory* (Vol. 150). New York: Springer.

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.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*(3), 237-247.

Stone, C. A., & Yeh, C.-C. (2006). Assessing the dimensionality and factor structure of multiple-choice exams: An empirical comparison of methods using the Multistate Bar Examination. *Educational and Psychological Measurement, 66*(2), 193-214.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*(3), 393-408.

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.

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

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: Theory and applications. *Applied Psychological Measurement, 26*(1), 109-128.

Wang, W. C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, *29*(2), 126-149.

Wilson, M., & Adams, R. J. (1995). Rasch models for item bundles. *Psychometrika*, *60*(2), 181-198.

Wilson, D. T., Wood, R., & Gibbons, R. (1998). *TESTFACT: Test scoring, item statistics, and item factor analysis.* Mooresville, IN: Scientific Software International.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: current approaches and future directions. *Psychological methods*, *12*(1), 58-79.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). *ConQuest: Generalized item response modeling software [Computer software and manual]*. Camberwell, Australia: Australian Council for Educational Research.

Yang-Wallentin, F., Jöreskog, K. G., & Luo, H. (2010). Confirmatory factor analysis of ordinal variables with misspecified models. *Structural Equation Modeling, 17*(3), 392–423.

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.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of educational measurement, 30*(3), 187-213.

Zhang, B. (2010). Assessing the accuracy and consistency of language proficiency classification under competing measurement models. *Language Testing, 27*(1), 119-140.

Zhang, B., & Stone, C. (2004, April). *Direct and indirect estimation of three-parameter compensatory multidimensional item response models*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.