The Bayesian Covariance Structure Model for Testlets

Jean-Paul Fox

Jeremias Wenzel

Konrad Klotzke

Department of Research Methodology, Measurement and Data Analysis,

University of Twente, the Netherlands

Corresponding author:
Jean-Paul Fox
Department of Research Methodology, Measurement and Data Analysis
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands
Phone: +3153 489 3326
E-mail: j.p.fox@utwente.nl

The Bayesian Covariance Structure Model for Testlets

May 2020

Abstract

Standard item response theory (IRT) models have been extended with testlet effects to account for the nesting of items; these are well known as (Bayesian) testlet models or random effect models for testlets. The testlet modeling framework has several disadvantages. A sufficient number of testlet items are needed to estimate testlet effects, and a sufficient number of individuals are needed to estimate testlet variance. The prior for the testlet variance parameter can only represent a positive association among testlet items. The inclusion of testlet parameters significantly increases the number of model parameters, which can lead to computational problems.

To avoid these problems, a Bayesian covariance structure model (BCSM) for testlets is proposed, where standard IRT models are extended with a covariance structure model to account for dependences among testlet items. In the BCSM, the dependence among testlet items is modeled without using testlet effects. This approach does not imply any sample size restrictions and is very efficient in terms of the number of parameters needed to describe testlet dependences. The BCSM is compared to the well-known Bayesian random effects model for testlets using a simulation study. Specifically for testlets with a few items, a small number of test takers, or weak associations among testlet items, the BCSM shows more accurate estimation results than the random effects model.

Introduction

Many tests have sections consisting of sets of items that are related to a common stimulus (e.g., reading passage, data display). Each set of items is referred to as a testlet, and it is well known that the relationship between the items and the common stimulus can lead to positive dependence among the item responses of an individual (Wainer, Bradlow, & Wang, 2007). Given the test taker's ability level, the item responses within each testlet are positively correlated, which leads to a violation of the local independence assumption of IRT models. This positive dependence structure cannot be ignored. When item responses are incorrectly assumed to be (conditionally) independent, the precision of the ability estimates will be overestimated and the ability and item parameter estimates will contain bias (Wainer, Bradlow, & Du, 2000). Conversely, when item responses are incorrectly assumed to be dependent, the precision of the ability estimates can be underestimated and item parameter estimates can contain bias. In general, meaningful statistical inferences with an IRT model require careful handling of the dependence structure. This has led to much discussion in the literature on reliable methods to evaluate the assumption of local independence in IRT models.

In previous research, it has been shown that when tests are constructed with testlet items, the items appear to be dependent even after conditioning on the latent variable (Bradlow, Wainer, & Wang, 1999; Li, Bolt, & Fu, 2006; Sireci, Thissen, & Wainer, 1991; Thissen, Steinberg, & Mooney, 1989; Wainer & Kiely, 1987; Wang, Bradlow, & Wainer, 2002). The item responses tend to be more related to each other than can be explained by the (unidimensional) latent variable. The level of dependence among the testlet items depends on the level of the testlet variance, (i.e., the variance of the testlet effects; see below). Items can be assumed to be locally independent when the testlet variance is equal to zero. The greater the variance, the larger the dependence among the items.

A popular model for testlets is the Bayesian random effects model (Bradlow et al., 1999; Wainer et al., 2000), referred to as a testlet response theory (TRT) model. This modeling approach includes a random effect to capture the dependence among item responses within a testlet. This random effect approach has several limitations. Testlet designs are easily incorporated in a test, but the random effect models for testlets (e.g., Bradlow et al., 1999) are subject to strict sample size restrictions. Each testlet needs to consist of a sufficient number of items, and the testlet needs to be administered to a sufficient number of test takers. This limits the applicability of testlet designs in practice. When the design is incomplete, for instance due to the adaptive nature of the test, these sample size restrictions can become a problem. The assumption of local independence is a pressing issue that needs to be addressed in smaller-scale applications.

For the TRT model, the number of testlet parameters can easily become overwhelming. A testlet parameter is introduced to address the dependence for each combination of testlet and test taker. When the test contains many testlets and is administered to a large number of test takers, a huge number of parameters are needed to model the testlet dependence structure. This is a very inefficient parameterization, since the strength of dependence depends only on the testlet variance parameter, which is assumed to be the same across individuals. In the proposed model, each testlet dependence can be modeled with a single covariance parameter, without the need to include a testlet effect for each test taker. In practice, interest usually isn’t focused on the test taker’s testlet effects. Testlet effects are only used to model the dependence structure. However, they can seriously complicate the computational burden and imply restrictions on the sample size. They also complicate the interpretation of estimated item parameters, since their scale depends on the testlet effect parameters and the trait parameter.

Another issue with the TRT model is the prior distribution for the testlet variance. The testlet variance parameter determines the strength between testlet items and is restricted to be positive. A noninformative prior for the variance parameter is a topic of much discussion. When the testlet variance is close to zero, the inverse gamma prior can introduce a bias by overstating the level of dependence of the testlet items. Furthermore, the inverse gamma prior is unable to include the point of no testlet variance, representing the state of an independent set of items. This makes it difficult to verify whether a set of items is nested within a testlet. The prior information for the level of dependence excludes the option that the items do not correlate. Furthermore, a testlet variance of zero is of specific interest, but this point lies on the boundary of the parameter space. Classical test procedures such as the likelihood ratio test can break down and have complex sampling distributions, which complicates the computation of critical values, when the true parameter value is on the boundary.

A Bayesian covariance structure model (BCSM; Klotzke & Fox, 2019a, 2019b) for testlets is proposed to address the shortcomings of the TRT model. The BCSM also modifies the standard IRT models to accommodate the clustering of items: The covariance structure of the errors is modeled to handle the greater dependence of items within testlets. In this additional covariance structure, dependences between clusters of items are modeled. This parameterization is very efficient, since a common covariance can be assumed between responses to items in a testlet across test takers. Therefore, the number of additional model parameters for the BCSM is equal to the number of testlets, when assuming testlet-specific dependences. The prior information for the level of dependence can also include no dependence between items in a testlet. Furthermore, testlet effects do not have to be estimated, since they are not needed to model the dependences. For the BCSM, sample size restrictions can be relaxed in comparison to the TRT model, which makes the BCSM more suitable for small sample sizes, incomplete designs, and extensive testlet structures with many testlets, each containing just a few items.

In the remainder of this report, the TRT model is described as a modification of standard IRT models for binary and polytomous data. Then, the BCSM for testlets is presented, where covariance structure models are discussed as extensions of IRT models. The computational method to estimate the parameters is based on Markov chain Monte Carlo (MCMC), which is briefly described. Then, simulation results for several test designs are presented, which include a comparison between the TRT model and the BCSM for testlets. In the final section, conclusions and model generalizations are discussed.

Testlet Response Theory

TRT models were introduced as a model-based approach to handle violations of local independence, since sets of items are related to a single stimulus (Bradlow et. al., 1999; Wainer, et al., 2000; Wang & Wilson, 2005). In TRT, a standard unidimensional IRT model is extended to include testlet parameters to account for within-testlet dependence. TRT models can be viewed as a confirmatory multidimensional IRT (MIRT) model in which all item responses are influenced by a common latent trait, and item responses within a testlet are further explained by a testlet parameter.

In this research, the considered data structure consists of N examinees $(i = 1,\ldots,N)$ who receive a test of K items ($k = 1,\ldots,K$), which are scored in a binary or polytomous fashion. A completely observed data matrix (Y) is assumed, where sets of items are clustered in testlets. A total of D testlets ($d = 1,\ldots,D$) is assumed, and item k is assigned to testlet $d(k)$, where $d(k)$ represents the testlet that contains item k. The size of each testlet is represented by n_d. When the number of testlets is equal to the number of items, each item is in its own testlet. Items cannot be assigned to more than one testlet.

A Probit version of the two-parameter TRT model is considered, where the probability of a correct response of test taker i to item k, assigned to testlet $d(k)$, is represented by

$$P\left(Y_{ik} = 1 \middle| a_k, b_k, \gamma_{id(k)}\right) = \Phi\left(a_k(\theta_i - \gamma_{id(k)}) - b_k\right), \tag{1}$$

$$\begin{array}{lcl} \theta_i & \sim & N\left(0, \sigma_\theta\right) \\ \gamma_{id(k)} & \sim & N(0, \sigma_{\gamma_d}) \end{array}$$

where $\gamma_{id(k)}$ represents the testlet effect of item k to test taker i, with item k nested in testlet $d(k)$. Therefore, extra dependence of items within a testlet is modelled with the random effect $\gamma_{id(k)}$. The testlet variance parameter σ_{γ_d} represents the variance in testlet effects across test takers and can be assumed to be testlet specific. The shared testlet effect for items within a testlet gives rise to a correlation between item responses of a person for that testlet. The testlet effect adds a negative (positive) contribution to the success of a correct response when $\gamma_{id(k)} > 0$ ($\gamma_{id(k)} < 0$). The probability of a correct response is reduced (increased) when the item is nested in a testlet.

The model specification is completed with a prior specification for the parameters. Normal distributions are assumed for the item parameters, where the priors are given by:

$$\begin{aligned} a_k &\sim N\left(\mu_a, \sigma_a^2\right) \\ b_k &\sim N\left(\mu_b, \sigma_b^2\right). \end{aligned}$$

Inverse gamma priors are specified for the variance parameters $(\sigma_{\gamma_d}, \sigma_a^2, \sigma_b^2)$, with g_1 and g_2 being the shape and scale parameters, respectively. The prior is assumed to be vague, with shape and scale parameters equal to 0.01. Finally, the mean parameters μ_a and μ_b have noninformative priors ($p(\mu_a) \propto c,\ \ p(\mu_b) \propto c,\ c > 0$).

The testlet effects are zero-centered to identify the model and to interpret the testlet effects as deviations from the standard linear predictor in the two-parameter IRT model. Furthermore, the mean and variance of the ability distribution are set to zero and one, respectively, to identify the model.

The BCSM for Testlets

The BCSM also modifies the standard IRT models by including a covariance structure model for the extra dependence of items within a testlet. When representing the two-parameter IRT model in a latent variable form, where latent responses Z_{ik} underlie the observed binary response Y_{ik}, a normally distributed error term is introduced. This error term represents the randomness in a response across hypothetical replications of the item to a test taker. Then, the two-parameter IRT model is represented by

$$\begin{array}{lcl} Z_{ik} & = & a_k\theta_i - b_k + e_{ik} \\ \theta_i & \sim & N(0,\sigma_\theta) \\ e_{ik} & \sim & N(0,1), \end{array} \tag{2}$$

where $Y_{ik} = 1$ if $Z_{ik} > 0$ and $Y_{ik} = 0$ if $Z_{ik} \leq 0$. The responses of test taker i are assumed to be independently distributed, since the errors are independently distributed.

When sets of items are nested in testlets, the errors are assumed to be dependent within each testlet. Consider the responses to items in testlet d of individual i, $\mathbf{Z}_{id}$. They are assumed to be multivariate normally distributed to model the dependence between items in a testlet. The covariance matrix of the errors represents the dependence due to the nesting of items in testlet d. To illustrate this, assume that the first two items are nested in testlet $d = 1$. A multivariate two-parameter IRT model is defined by assuming a multivariate distribution for the error distribution in Equation (2). Then, the responses (Z_{i1}, Z_{i2}) are multivariate normally distributed,

$$\begin{array}{lcl} \left(Z_{i1}, Z_{i2}\right) & = & \left(a_1\theta_i - b_1, a_2\theta_i - b_2\right) + \left(e_{i1}, e_{i2}\right) \\ \left(e_{i1}, e_{i2}\right) & \sim & N\left(\mathbf{0}, \mathbf{\Sigma}_d\right), \end{array}$$

and $\mathbf{\Sigma}_d$ has diagonal elements $1+\sigma_{\gamma_1}$ and non-diagonal components σ_{γ_1}. The covariance parameter represents the common covariance between responses to items within a testlet. The level of dependence is represented by covariance parameter σ_{γ_1}, which corresponds to the testlet variance parameter in the TRT model. It follows that the variance parameter describes the dependence among testlet items in the TRT model, which is restricted to be positive. So, under the TRT model, it is not possible to describe negative associations among testlet items. For instance, in a testlet with technology-enhanced items, the success probability can increase by heightening the engagement of a test taker. At the same time, multiple sources of information (e.g., paired text passages) can complicate testlet items, leading to a reduced success probability for items in the testlet (e.g., Jiao, Lissitz, & Zhan, 2017). A negative correlation can occur among responses to testlet items, since the innovative character of the testlet can stimulate the success probabilities positively for some testlet items but negatively for others. In general, the complex nature of innovative testlet items can lead to negative associations among responses to the testlet items due to the diverse effect of the testlet on the success probabilities.

The covariance structure $\mathbf{\Sigma}_d$ follows from the TRT model when the discrimination parameters are equal to one. Consider the latent response model for the TRT $Z_{ik} = \theta_i - b_k - \gamma_{id(k)} + e_{ik}$, and consider the term $t_{ik} = \gamma_{id(k)} + e_{ik}$ as the error component of the model. This error component consists of two normally distributed variables, and thus t_{ik} is also normally distributed. The mean is zero and the variance of the t_{ik} equals the sum of the variances $\sigma_{\gamma_d} + 1$. The covariance of responses to items k and k' in testlet d of individual i is equal to the covariance of t_{ik} and $t_{ik'}$, which is equal to the variance σ_{γ_d}. It follows that the BCSM for testlets directly models the extra covariance among responses to testlet items, where the TRT model includes a random effect to model this additional dependence.

In this research, the dependence structure of the BCSM does not include the item discrimination parameters. In the TRT model in Equation (1), the discrimination parameters also influence the dependence structure through multiplication with the testlet effects. In the BCSM, a homogeneous association among items in a testlet is assumed. The inclusion of discrimination parameters is described in the discussion, which is a topic of further research.

The BCSM for testlets can be represented as a multivariate distribution for the responses to items in a testlet d. The superscript d is used to represent a vector of

parameters or a vector of observations for testlet d. For instance, the latent responses to testlet d of individual i is represented by $\mathbf{Z}_i^d$. The latent variable form is used, and the latent responses to items in testlet d are assumed to be
multivariate normally distributed,

$$\begin{array}{lcl} \boldsymbol{Z}_i^d & = & \boldsymbol{a}^d\theta_i - \boldsymbol{b}^d + \boldsymbol{e}_i^d \\ \theta_i & \sim & N(0,\sigma_\theta) \\ \boldsymbol{e}_i^d & \sim & N(0,\boldsymbol{\Sigma}_d), \end{array} \tag{3}$$

where $\boldsymbol{\Sigma}_d = \mathbf{I}_{n_d} + \mathbf{J}_{n_d}\sigma_{\gamma_d}$, and where $\mathbf{I}_{n_d}$ and $\mathbf{J}_{n_d}$ are the identity matrix and a matrix of ones, respectively, both of dimension n_d.

The main difference between the TRT model (Equation [1]) and the BCSM (Equation [3]) is that the TRT model has a testlet effect parameter to model the dependence structure, where the BCSM describes the extra dependence with a covariance matrix. The BCSM does not include any testlet parameters, which leads to a serious reduction in the number of model parameters. The BCSM is much more efficient in describing the dependence structure. Furthermore, σ_{γ_d} is a covariance parameter in the BCSM, which can also be negative or zero. This is in contrast to the TRT model, where σ_{γ_d} is a variance parameter, which is restricted to be positive. However, the covariance matrix $\boldsymbol{\Sigma}_d$ must be positive definite, which restricts the covariance parameter σ_{γ_d} to be greater than $-1/n_d$, where n_d is the number of items in testlet d (i.e., the dimensionality of the covariance matrix).

The extension to polytomous response data is straightforward. In the latent response formulation, the latent responses are assumed to be truncated multivariate normally distributed. For an observed response in category c, the corresponding latent response is restricted to be greater than the upper bound for category $c-1$ and less than the lower bound for category $c+1$. For ordinal response data, the category boundaries follow an order restriction. The latent response formulation for polytomous IRT models can be found in Fox (2010).

Bayesian Inference

An MCMC method is used to draw samples from the posterior distributions of the model parameters to make inferences about the unknown model parameters. For the binary TRT model, this method is described in Bradlow et al. (1999). The authors implemented a Gibbs sampler to draw parameter values from their conditional distributions. Wang et al. (2002) proposed Metropolis–Hastings steps to make draws from the conditional distributions for parameters of the polytomous TRT model for ordinal data. The R-package sirt (Robitzsch, 2019) contains MCMC algorithms for binary and polytomous TRT models.

For the BCSM, an MCMC algorithm is proposed (a full description can be found in the Appendix). The novel steps of the algorithm are explained in more detail. This includes the conditional distribution of the parameter σ_{γ_d}, which requires specific attention. First, the prior specification of the model parameters is discussed. The priors for the item and ability parameters are similar to those in the TRT model. However, the prior for σ_{γ_d} is different, since this is a covariance parameter in the BCSM. The technique of Fox, Mulder, and Sinharay (2017) is used to obtain the posterior distribution of σ_{γ_d}. Consider the distribution of the variable $\tilde{Z}_{id} = \sum_{k \in d} \tilde{Z}_{id(k)} / n_d$, where $\tilde{Z}_{id(k)} = Z_{id(k)} - (a_k \theta_i - b_k)$. This variable is normally distributed with mean zero and variance

$$
\begin{aligned}
Var\left(\tilde{Z}_{id}\right) &= \frac{1}{n_d^2}\left(\sum_{k\in d} Var\left(\tilde{Z}_{id(k)}\right)+\sum_{k,k'\in d} Cov\left(\tilde{Z}_{id(k)},\tilde{Z}_{id(k')}\right)\right)\\
&= \left(1+\sigma_{\gamma_d}\right)/n_d+(n_d-1)\sigma_{\gamma_d}/n_d\\
&= 1/n_d+\sigma_{\gamma_d}.
\end{aligned}
$$

The posterior distribution of σ_{γ_d} is constructed from the normally distributed $\boldsymbol{Z}_d$ and the prior for σ_{γ_d}; that is,

$$
p\left(\sigma_{\gamma_d} \mid \boldsymbol{Z}_d,\boldsymbol{\theta},\boldsymbol{a}^d,\boldsymbol{b}^d\right) = \left(2\pi(1/n_d+\sigma_{\gamma_d})\right)^{-N/2}\exp\left(\frac{-\sum_i \tilde{Z}_{id}^2/2}{1/n_d+\sigma_{\gamma_d}}\right)p\left(\sigma_{\gamma_d}\right).
$$

A conjugate prior for σ_{γ_d} is defined, which is known as a shifted inverse gamma distribution (Fox et al., 2017). This prior distribution can be represented as

$$
\text{shifted-IG}\left(\sigma_{\gamma_d};g_1,g_2,1/n_d\right) = \frac{g_2^{g_1}}{\Gamma(g_1)}\left(\sigma_{\gamma_d}+1/n_d\right)^{-g_1-1}\exp\left(-\frac{g_2}{\sigma_{\gamma_d}+1/n_d}\right),
$$

where g_1 is the shape parameter, g_2 is the scale parameter, and $1/n_d$ is the shift parameter. As a result, the posterior distribution of σ_{γ_d} is also shifted inverse gamma distributed, with shape parameter $N+g_1$, scale parameter $g_2+\sum_i \tilde{Z}_{id}^2/2$, and shift parameter $1/n_d$. From the shift parameter it follows that the covariance parameter σ_{γ_d} is restricted to be greater than $-1/n_d$. Samples can be drawn directly from the shifted inverse gamma posterior. Therefore, we sampled values for $\sigma_{\gamma_d}+1/n_d$ from the inverse gamma distribution and subtracted $1/n_d$ from the sampled values.

The conditional distributions of the remaining parameters for the binary BCSM are described in the Appendix. For polytomous data, the conditional distributions of the BCSM parameters follow in a similar way. Although the testlet effects are not included in the BCSM, testlet effects can still be estimated under the BCSM by extracting them from the residuals.

This is only possible when the covariance parameter is positive, and a testlet effect can describe the positive associations among the testlet residuals. In that case, we use the model description of the TRT model, $Z_{ik} = (a_k\theta_i - \gamma_{id(k)}) - b_k + e_{ik}$, without multiplying the testlet parameter by the discrimination parameter. The result is that the conditional distribution of the testlet parameters γ_{id} is normal with mean

$$E\left(\gamma_{id} \mid \mathbf{Z}_i^d, \theta_i, \mathbf{a}^d, \mathbf{b}^d, \sigma_{\gamma_d}\right) = \frac{-\sum_{k \in d} Z_{ik} - (a_k\theta_i - b_k)}{1/\sigma_{\gamma_d} + n_d} \quad (4)$$

and variance $(1/\sigma_{\gamma_d} + n_d)^{-1}$. A post hoc sample can be obtained by drawing samples for the testlet parameter given the sampled BCSM parameters. Note that the covariance parameter σ_{γ_d} also contributes to the variance of the scale on which the testlet effects are estimated.

The BCSM is identified by restricting the mean and variance of the (primary) latent variable θ_i. This is done by restricting the mean of the latent variable to zero. The mean of the discrimination parameters is restricted to one to fix the variance of the scale.

Simulation Study: TRT Versus BCSM

The performance of the TRT model for testlets was examined in a simulation study and compared to the performance of the BCSM model using MCMC for parameter estimation. The first purpose of the simulation study was to compare the BCSM to a current standard approach in the test industry, specifically the TRT model. In this comparison, the testlet variance was also restricted to be small in order to examine the influence of the prior information under both models. The second purpose was to confirm that accurate parameter estimates can be obtained for the BCSM under a wide variation of experimental factors. Three factors were varied across the simulation conditions: (a) the number of examinees, (b) the number of items per testlet, and (c) the testlet variance. The parameter values were set to realistic values and defined similarly to the ones used by Wang et al. (2002).

In this comparison, data was simulated under the TRT model. A population distribution was defined for the model parameters in order to generate data under the TRT model. The simulated datasets mimic real-world applications, assuming the TRT model structure is true for a real population. In this simulation study, binary responses were simulated, where a one indicates success and a zero indicates no success. The datasets were used to estimate the parameters for both models. The following population distributions were asserted for the generation of datasets: $\theta_i \sim N(0,1)$ and $b_k \sim N(0,.5)$. The discrimination parameters were equal to one to facilitate a comparison between the TRT model and the BCSM. The discrimination parameters did not affect the covariance structure in the BCSM, when they are equal to one. The testlet variance is directly related to the variance of the ability parameter θ_i. A testlet variance of 0.5 would indicate that the testlet variance is half the size of the ability variance. The testlet effects were simulated from a normal distribution with a mean of zero and different testlet variances. Then, data was generated from the TRT model (Equation [1]) given the generated testlet effects, difficulty, and ability parameters.

Since the data was generated under the TRT model, it was expected that estimation results would be similar for moderate to large sample sizes. For small sample sizes, it was expected that the estimated testlet variance parameters under the TRT model would show bias, which was also observed by Jiao, Wang, and He (2013). The prior for the testlet variance can lead to an overestimation when the testlet effect is small.

The TRT model is only statistically equivalent to the BCSM for a fixed positive (co)variance parameter σ_{γ_d}. When considering σ_{γ_d} to be a random variable, the posterior distribution of σ_{γ_d} is restricted to a positive parameter space under the TRT model. Under the BCSM, the σ_{γ_d} can also take on negative values, which means that the items within a testlet can correlate negatively. The parameter space is less restricted under the BCSM, and for that reason the posterior standard deviation is expected to be slightly higher in comparison to that under the TRT model. From this perspective, the simulation study favored the TRT results, since the data was generated under the TRT model. Nevertheless, the simulation results showed a better performance of the BCSM than of the TRT model.

Each condition was replicated 1,000 times. The number of items was set to $K = 30$ for each condition. The number of participants was set to 1,000, 500, and 200. Second, the number of items per testlet was varied to 5, 10, and 15 items per testlet, corresponding to 6, 3, and 2 testlets per dataset, respectively. Finally, the variance of the testlet effect was manipulated. Small testlet variances tend to be more common in practice and thus were the focus of this study. Therefore, the choice was made to use a testlet variance of $\sigma_d^2 = .1,\ .05,\ \text{and}\ .01.$ A Latin-square design was used to handle the variation of the three experimental factors. In a full factorial design, 27 conditions would have to be evaluated, which would have required extensive computation time. For relatively large sample sizes, significant differences were not expected to be found between the BCSM and the TRT model given the simulation results of Wang et al.

(2002) and Bradlow et al. (1999), which would make a full factorial design very inefficient. Instead of varying the third factor, representing the testlet variance, over all possible combinations of the first two factors, the third factor was used only once for every factor. The resulting simulation design had nine distinct conditions. In Table 1, the whole design is shown, where the numbers 1 through 9 in parentheses refer to the numbering of the nine experimental conditions.

TABLE 1

Table of simulation design

		# Items per Testlet		
Variance of the Testlet Effect		5	10	15
No. of participants	1,000	.05 (1)	.1 (2)	.01 (3)
	500	.1 (4)	.01 (5)	.05 (6)
	200	.01 (7)	.05 (8)	.1 (9)

MCMC procedures were used to estimate the model parameters of the TRT model and the BCSM. When convergence was reached, additional samples from the posterior distribution were drawn to make inferences. The advantage of this is that further inferences, like computing the mean of the posterior distribution of a parameter, can be easily made. In this study, the first 1,000 iterations were discarded as the burn-in, and another 9,000 iterations were made to the model parameters. The Heidelberger and Welch's convergence diagnostic (Plummer, Best, Cowles, & Vines, 2006) was used to evaluate the convergence of the MCMC chains. The average lag-50 autocorrelation and the effective sample size were computed to compare the estimation performance of both MCMC algorithms. For M independent MCMC draws, the Central Limits Theorem states that the bound on the estimation error is proportional to $1/\sqrt{M}$. For M dependent MCMC draws, the bound is proportional to

$1/\sqrt{M_{eff}}$, with M_{eff} the effective sample size. The M_{eff} represents the number of independent values that has the same estimation performance as *M* MCMC correlated values in the MCMC chain.

The estimated model parameters were compared to the true simulated values. The posterior means were used as parameter estimates. Three criteria were used to assess the quality of the parameter estimates. The bias, root mean squared error (RMSE), and coverage rate were computed for each parameter. For instance, for the item difficulty parameter $b_k,$ the bias was computed as the average over $R=1000$ replications:

$$Bias(\hat{b}_j) \quad = \quad \frac{1}{R}\sum_{r=1}^{R}(\hat{b}_{jr}-b_{jr}), \tag{5}$$

where $\hat{b}_{kr}$ is the estimated difficulty parameter for item k and b_{kr} the true value in replication $r.$ The RMSE was calculated as the square root of the MSE,

$$MSE(\hat{b}_j) \quad = \quad \frac{1}{R}\sum_{r=1}^{R}\left(\hat{b}_{jr}-b_{jr}\right)^2. \tag{6}$$

The bias and RMSE were calculated for different model parameters obtained under the two models. Where applicable, the results were averaged across the test items and participants in order to make the comparison more straightforward.

A 95% highest posterior density (HPD) interval was used to compute the coverage rates. The coverage rate gives an indication of the number of times that the true value lies within the 95% HPD interval. The data was simulated under the TRT model, so for the TRT model the coverage rates were designed to be 95%. When the coverage rate is lower, the true parameter value is less often recovered than would be expected according to the 95% HPD interval. A low coverage rate indicates a problem with the model’s functioning. The coverage rate under the BCSM was expected to be higher than 95%, since data was simulated under the TRT model. Under the BCSM, the 95% HPD intervals are wider, since the testlet correlations are not restricted to be positive. As the data was generated under the TRT model, the coverage rates under the BCSM were therefore not expected to match up exactly.

Results

The MCMC algorithm for the BCSM model performed good. Several problems were found with the estimation performance of the MCMC algorithm for the TRT model. For moderate to large sample sizes and a testlet variance of .1, the MCMC samples under the TRT showed reasonable stable behavior and better convergence. For smaller testlet variances, the chains under the TRT model showed much higher autocorrelations and less smooth transitions through the parameter space than the chains under the BCSM (Figure 1). For the TRT model, the subplots show the trace plots of the sampled testlet variances for *N* = 200 and *N* = 1,000. For *N* = 200, the chain shows high correlation between sampled values. When values close to zero were sampled, the chain showed difficulties in moving away from the state of no testlet variance. This can be seen around iteration numbers 4,000 and 8,000. The MCMC chain shows problems in moving away from the state of no testlet variance. For *N* = 1,000 and a higher true testlet variance, this problem did not occur. However, the trace plots still show highly correlated values. For the BCSM, the behavior of the MCMC chain is much better. For small sample sizes, the chain did not got stuck at zero, since negative testlet covariances were allowed. Without this lower bound at zero, the movement of the chain through the entire parameter space was improved, leading to more informative MCMC samples (less correlated samples) from the posterior distribution. When increasing the sample size to N=1000, the behavior of the chain under the BCSM remained almost similar to the one for N=200.

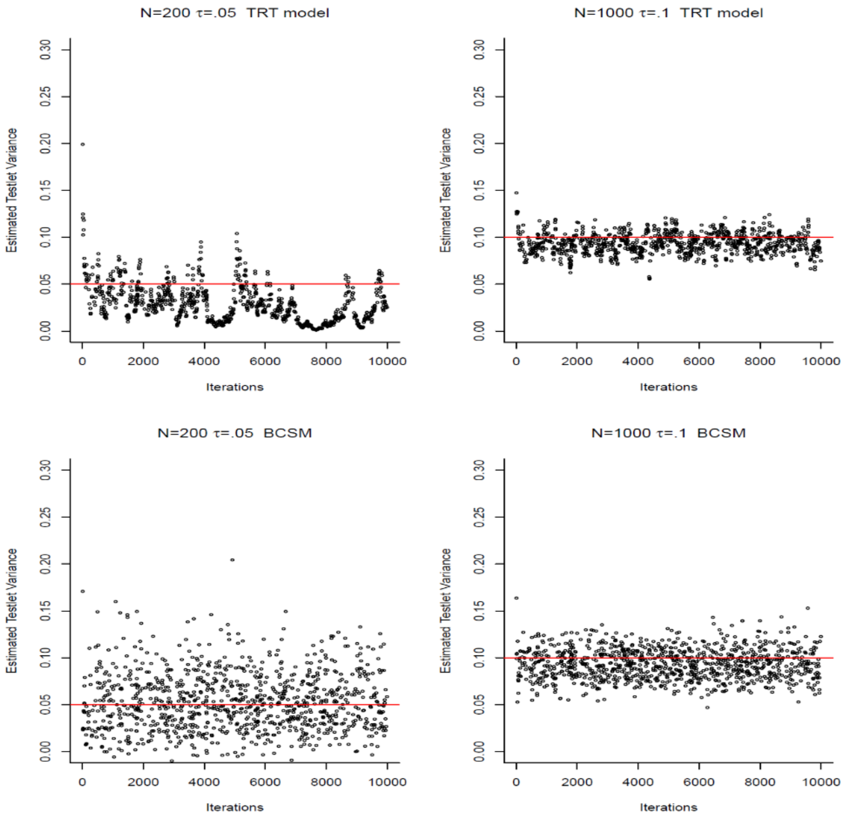

FIGURE 1. *Trace plots of MCMC chains for small and moderate sample sizes under the TRT model and the BCSM*

For each condition, the average effective sample size and lag-50 autocorrelation was computed for the testlet variance across replications under each model. For the TRT model, it can be seen in Table 2 that for 9,000 MCMC iterations the M_{eff} is dramatically low in all conditions. The low M_{eff} corresponds to high lag-50 autocorrelations. The effective sample sizes under the TRT model are much lower than those computed under the BCSM. The effective sample sizes under the BCSM are around 8%-18% of the total number of MCMC iterations (after the burn-in period). The MCMC samples under the BCSM provide acceptable bounds on the estimation error. The estimated average lag-50 autocorrelation is also close to zero for the BCSM chains. To obtain the same estimation performance with the TRT, in theory 10 to 50 times more MCMC iterations are needed leading to 100,000 to 500,000 MCMC iterations. Then, the simulation study would take too much time to be completed. However, the main problem was that the MCMC chains got stuck at zero, and increasing the number of MCMC iterations did not solve that problem. It led to more highly autocorrelated samples.

TABLE 2

The (MCMC) effective sample size (M_{eff}) and lag-50 autocorrelation averaged across replications per condition and model.

	TRT		BCSM	
Condition #	M_{eff}	Lag 50	M_{eff}	Lag 50
1	67	.481	1523	-.001
2	203	.116	1731	-.001
3	20	.810	743	.000
4	85	.422	1748	.000
5	18	.820	1032	-.002
6	100	.384	1063	-.001
7	24	.773	1278	-.001
8	55	.603	1367	-.001
9	200	.161	1287	.000

For the 9 conditions, the estimates for the discrimination, difficulty, and ability parameters were obtained. The true parameters for a, b, and θ were accurately recovered. No apparent differences were found between the parameter estimates under the two models. Estimates for bias and RMSE of these parameters showed an accurate recovery of the true values; no significant differences were found in terms of bias and RMSE between the two models.

The estimated testlet variance parameters under both models are shown in Table 3. We see that under the TRT model, the true testlet variance was generally underestimated. The posterior mean was used as an estimator but the estimation results did not differ much from those using the posterior mode as an estimator. A vague inverse gamma prior for the variance parameter was used, with the shape and scale parameter equal to .01. For a true testlet variance value of .01, the testlet variance was sometimes not detected under the TRT model. Specifically in Condition 7, a testlet variance was not detected in many of the replications (95% coverage rate equaled .13) and the estimated testlet variance was around .0038 under the TRT model. For this condition, it was investigated if the testlet variance estimates improved under the TRT model, when using an inverse gamma prior with a shape and scale value equal to one. This prior gave more support to higher testlet variances. The estimated testlet variance for 1,000 replications equaled .052, which is more than 13 times greater than the estimate with the vague prior with shape and scale parameters equal to .01, and 5 times greater than the true value of .01. The RMSE of the estimated testlet variance was .051 and the coverage rate was equal to zero. In conclusion, the prior can be adjusted to cover higher testlet variances, but this easily leads to overestimating the true value. These estimation problems did not occur under the BCSM, where a vague prior was used for all conditions, and all testlet variances were accurately estimated.

TABLE 3

Estimated testlet variances of the TRT model and the BCSM across 1,000 data replications

Condition #	σ_γ	$\hat{\sigma}_\gamma$	
		TRT	BCSM
1	.05	.0516	.0503
2	.10	.0734	.1007
3	.01	.0065	.0109
4	.10	.0790	.1019
5	.01	.0063	.0100
6	.05	.0303	.0520
7	.01	.0038	.0110
8	.05	.0266	.0544
9	.10	.0628	.1092

In Table 4, the RMSEs of the testlet variances and testlet effects under both models are presented. The estimated differences in RMSE of the testlet variance parameters were very small, and the RMSEs were overall very small. Estimation of the RMSE and bias was based on mean parameter estimates. It was expected that under the TRT model, a skewed posterior distribution of the testlet variance parameter would lead to an overestimation of the true parameter value. Significant overestimation of the true value due to skewed posteriors was not detected. Estimates of testlet effects under the BCSM were comparable to those under the TRT model, when they were transformed to a common scale with mean zero and an equal testlet variance. Estimated testlet effects under the BCSM were less shrunken toward the prior mean. This led to a greater number of outliers and more variance in estimated testlet effects.

TABLE 4

RMSEs of true parameters and estimated posterior mean for the testlet variance and testlet effects.

Condition #	Testlet Variance	Testlet Effect

	TRT	BCSM	TRT	BCSM
1	.0141	<.0100	.2663	.2665
2	<.0100	<.0100	.3294	.3294
3	<.0100	<.0100	.1277	.1284
4	.0224	.0173	.3491	.3491
5	<.0100	<.0100	.1285	.1292
6	.0141	.0141	.2512	.2516
7	<.0100	.0224	.1315	.1319
8	.0283	.0224	.2546	.2550
9	.0300	.0316	.3289	.3298

Note: The MSE was calculated with a precision of four decimals. When the MSE was less than .0001, the RMSE was less than .0100.

The 95% coverage rates are presented in Table 5. A remarkable finding was that the coverage rates for the TRT model were not satisfactory for most conditions (i.e., they were generally too low). In Condition 2, for a large sample size with higher testlet effects, the coverage rate was close to the target 95%, with a coverage of 93%. In many replications the true value was not captured by the 95% HPD interval. For small sample sizes and small testlet effects, the coverage rates were particularly low under the TRT model.

TABLE 5

The 95% coverage rates for the testlet variance under the TRT model and the BCSM

Condition #	TRT	BCSM
1	.85	1.00
2	.93	1.00
3	.87	.98
4	.87	1.00

5	.77	.99
6	.86	.98
7	.13	1.00
8	.69	1.00
9	.87	.97

Coverage rates for the BCSM were too high, as expected, and most often were 100%; parameter space for the testlet covariance parameter was also wider under the BCSM than under the TRT model. This leads to a wider posterior distribution of the testlet covariance parameter. As a result, the 95% HPD intervals are wider and the coverage rates higher under the BCSM than under the TRT model. Under the BCSM, the coverage rates were comparable across conditions, and did not show a relationship with sample size. In Conditions 6 and 9, the coverage rates under the BCSM were slightly smaller (98% and 97%, respectively).

Figure 2 shows the estimated posterior densities of the testlet variance parameter using an informative and a vague prior under the TRT model, given sampled values in Condition 7 (200 test takers and 5 items per testlet), where the true variance is .01. When using a vague prior, the posterior distribution is highly peaked at zero. In that case, the MCMC chain is stuck at zero, as shown in the upper subplot of Figure 1. When increasing the prior information about the testlet variance (i.e., shape and scale parameter of the inverse gamma is equal to one), the posterior distribution covers testlet variances above the true value. The 95% HPD interval is equal to $[.033, .087]$ and does not include the true parameter value. The prior gives less weight to small variance values and moves the posterior distribution toward higher variance values. This leads to an overestimation of the testlet variance and the small coverage rate in Condition 7, as shown in Table 5. Under the BCMS, the estimated posterior distribution of the covariance parameter σ_γ is centered around the true value and less skewed, since it also covers negative covariance values. The 95% HPD interval equals $[-.0737, .104]$. For the BCSM, the posterior distribution is not affected by the lower bound of zero, and valid draws from the posterior distribution are obtained even for a small sample size and a small true variance.

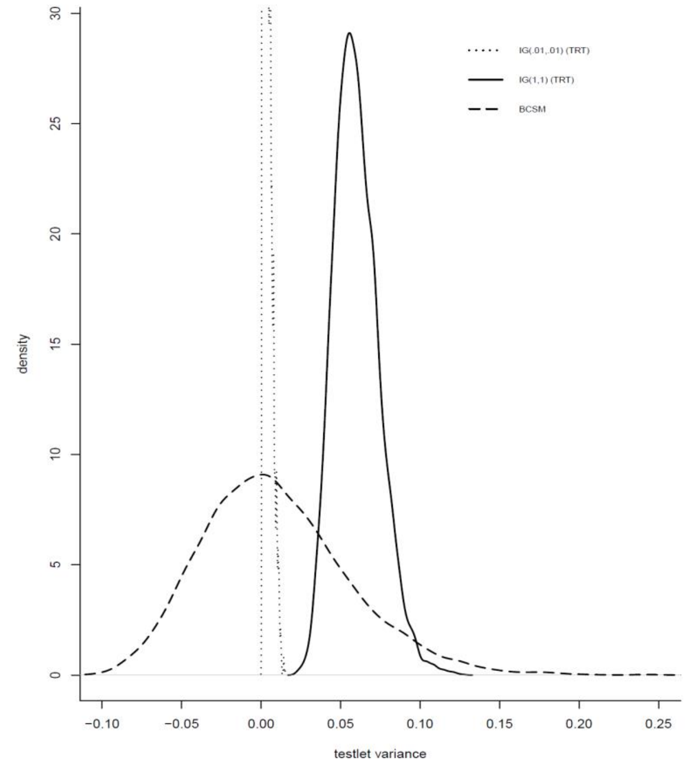

FIGURE 2. *Posterior density of the testlet (co)variance under the BCSM and the TRT model*

Polytomous Response Data

For the BCSM for polytomous data, the data was generated under the BCSM. The testlet covariance was varied across testlets, and ranged from −.05 to .50. The TRT model cannot handle a negative covariance among items in a testlet. In practice, negative associations among testlet responses can occur when the testlet leads to a stimulation of the success probability for some items but not others in the testlet. For instance, when a testlet consists of innovative items, directional local dependence and multidimensionality can lead to opposite stimulation of the success probabilities of the testlet items, which leads to a negative association.

Therefore, in this simulation study, the BCSM was not compared to the TRT model. Data was generated according to a BCSM for ordinal data, with three response categories, for a 10- and 20-item test. The number of test takers was equal to 1,000. The number of testlet items and the testlet variance were varied (Table 6). For the 10-item test, three testlets had 2, 3, and 5 items, respectively. For the 20-item test, three testlets had 4 items and one testlet had 8 items. In Table 6, the specific conditions are given under the label number of items, number of testlet items, and the testlet covariance parameter σ_γ.

TABLE 6

BCSM for polytomous data: Estimated testlet variance and 95% coverage probabilities

No. of Items	No. of Items Per Testlet	σ_γ	$\hat{\sigma}_\gamma$	*SD*	95% CR
10	2	−.05	−.042	.062	.929
	3	.10	.103	.054	.954
	5	.20	.205	.047	.950
20	4	−.05	−.048	.028	.948
	4	.10	.105	.036	.957
	4	.20	.205	.042	.941
	8	.50	.507	.047	.952

95% CR = 95% coverage rate

The following population distributions were asserted for the generation of the datasets: $\theta_i \sim N(0,1)$; thresholds were sampled from a uniform distribution between −1 and 2, and discrimination parameters were sampled from a log-normal distribution with a mean of zero and a standard deviation of .15. A total of 10,000 MCMC iterations were made for each data replication, where the first 1,000 were regarded as burn-in. A total of 1,000 data replications were made for each condition.

The posterior mean estimates of the testlet variances are given in Table 6, which are the average posterior means across replications. It can be seen that the estimated testlet variances are close to the true values, even for very small testlet variances. For the testlet with two items, the estimated negative covariance is slightly above the true value and the estimated coverage rate is around 93%. For a two-item testlet, it is more difficult to recover the negative correlation. It can be seen that the recovery is improved when the number of testlet items is increased to four. For the two-item testlet, the posterior standard deviation of the covariance parameter is relatively high, but is reduced by a factor of two when the number of testlet items is increased from two to four. The posterior standard deviation increases as values of the testlet covariance increase. In Table 5, the coverage probabilities under the BCSM were higher than the nominal level, since the data was generated under the TRT model. In this simulation study, the coverage probabilities are good and correspond closely to the nominal coverage probability. The other BCSM estimates (i.e., ability, discriminations, thresholds) were also close to their true values, and the corresponding coverage probabilities also matched the nominal level.

Discussion

The BCSM for testlets is a new modeling framework in which testlet dependences are modeled through an additional covariance structure. The BCSM has several advantages over TRT models. Under the BCSM, testlet effects do not have to be estimated, which leads to a serious reduction in the number of model parameters. In TRT models, for every combination of testlet and test taker a (random effect) parameter is introduced, which complicates estimation of the model parameters and comparison across different TRT models. Testlet effects also place restrictions on sample size: Each testlet needs to consist of a sufficient number of items to estimate the testlet effect. It was shown that for small sample sizes, it was difficult to specify the prior for the testlet variance. It was also shown that for the BCSM, accurate estimates were obtained for small sample sizes, small numbers of test takers (N = 200), and small numbers of items per testlet (2 items per testlet), without needing an informative prior.

Under the BCSM, testlet dependence is modeled with a covariance parameter. This avoids issues related to lower bound problems at zero when estimating model parameters. Testlet dependences are also allowed to be negative. The BCSM for polytomous testlet data also showed accurate estimation results. Furthermore, testlet effects can still be estimated under the BCSM through a post-hoc sampling approach. The issues in obtaining reliable posterior estimates under the TRT model were not clearly visible in terms of bias and RMSE of the ability and item parameters. Nevertheless, differences were notable in terms of the coverage rates and estimated testlet variances.

There have been various suggestions on how to handle testlet effects. Thissen, Steinberg, and Mooney (1989) suggested treating testlets as polytomous items and applying polytomous IRT models. However, this approach uses the same discrimination parameter for all items within a testlet and a total score for each testlet (Zenisky, Hambleton, & Sireci, 2002). These issues will possibly cause a loss of measurement information by having fewer parameters, and different scoring patterns for each testlet will be ignored. Wang and Wilson (2005) remarked that in a polytomous item approach, the number of response patterns of items within a testlet are modeled, and the testlet parameter is treated as fixed effects. For a testlet with 10 dichotomous items, this leads to 2^{10} testlet parameters. In the BCSM, the parameterization is much more efficient, where testlet dependences can be modeled with a single covariance parameter independent of the number of items within a testlet and the number of test takers.

In this research, the dependence structure of the BCSM did not include item discrimination parameters. The testlet effect parameter in the TRT model in Equation (1) is multiplied by a discrimination parameter. Then, the error term is $t_{ik} = a_k \gamma_{id(k)} + e_{ik}$. The dependence structure can be represented as the covariance between t_{ik} and $t_{ik'}$, which equals $a_k \sigma_{\gamma_d} a_{k'}$. So, for the TRT model in Equation (1), for each pair of items in a testlet the discrimination parameter also modifies the dependence structure. For the BCSM, this discrimination parameter is not included to model the dependence structure. Obtaining draws from the posterior distribution of the discrimination parameter is complicated when the discrimination parameter is included in the linear term and in the covariance matrix. To avoid this situation, it is possible to consider the error term of the linear predictor to be $t_{ik} = a_k(\theta_i - \gamma_{id(k)}) + e_{ik}$. The covariance matrix of the responses to testlet d of individual i can be represented as $\boldsymbol{a}_d \boldsymbol{a}_d^t \left(\sigma_\theta + \sigma_{\gamma_d}\right) + \boldsymbol{I}_d$, where the mean term of the linear predictors only contains the item difficulty parameters of the items in testlet d. In that case, the discrimination parameters are only included in the covariance matrix. The discrimination parameters can be included to modify the strength of the (positive or negative) testlet dependence among residuals represented by the testlet covariance parameter. More research is needed to obtain samples from their posterior distributions.

The TRT model is a restricted bi-factor model, where the discrimination parameters of the secondary factors are restricted to be proportional to the discriminations of the primary factor within each testlet (e.g., Rijmen, 2010). Although the bi-factor model is more flexible in describing the dependence structure of the testlet errors, the model has the same disadvantages as the TRT model. When assuming normally distributed factor variables, the model is identified by restricting the mean and variance of each factor to zero and one, respectively. Then, the linear function of the factor variables in the bi-factor model for an observations to item *k* in testlet *d* is represented by $t_{kd} = a_{kg}\theta_g + a_{kd}\theta_d + b_k + e_{kd}$, with multivariate normally distributed factor variables, $\boldsymbol{\theta} \sim N\left(\mathbf{0},\mathbf{I}\right)$. The covariance structure that includes the implied dependences by the secondary (testlet) factor is equal to $\boldsymbol{a}_d\boldsymbol{a}_d^t + \boldsymbol{I}_d$. It follows that the discrimination parameters $\boldsymbol{a}_d$ model the strength of dependence among testlet residuals, and they can be freely estimated. However, it is not possible to model a common negative dependence structure for a testlet with more than two items. Furthermore, for (very) small dependences among testlet residuals, the discrimination parameters are close to zero, which leads to problems in estimating the secondary factor θ_d. This secondary factor cannot be estimated, when the testlet discriminations parameters are equal to zero. This lowerbound problem is similar to the lowerbound issue of a zero factor variance in the TRT model. Finally, the sample size restrictions for parameter estimation for the TRT model also apply to the bi-factor model.

The BCSM differs from the well-known Gaussian copula, since it uses a structured covariance matrix to model the dependence structure. This structured covariance matrix follows from the joint conditional modeling approach of the marginal distributions. This dependence structure implied by the conditional model is integrated in the covariance structure of the BCSM and provides a clear interpretation of the parameters of the dependence structure. In copula modeling, the copula function determines the type of dependence and it operates directly on the marginals. For instance, in the Gaussian copula, marginal cumulative distributions functions are coupled using the multivariate normal cumulative distribution function with an unrestricted correlation matrix. This often leads to more complex dependence structures intertwining the factor and residual dependence.

Under the presented BCSM testlet model, and contrary to Gaussian copula models, closed-form expressions for the conditional posterior distributions of all model parameters are available. This allows to directly sample the parameters through an efficient Gibbs-sampling algorithm. Therefore, the BCSM does not require the numerical evaluation of integrals and offers an uncomplicated way to make inferences about the item parameters, person parameters, and the dependence structure.

The copula model preserves the marginal cumulative distributions in the construction of a multivariate distribution, which makes the modeling framework more flexible to define a multivariate distribution for any set of marginal cumulative distributions. However, estimating a fully parametric Gaussian copula model for categorical data is challenging and computationally expensive as it requires the evaluation of multivariate normal integrals in high dimensions (Pitt, Chan, & Kohn, 2006). A semi-parametric approach, on the other hand, neglects the information in the data about item and person parameters, and is therefore of limited utility in educational measurement applications (Hoff, 2007). The Gaussian copula is usually avoided for categorical data, since it requires intensive computation to evaluate the multivariate normal distribution (Braeken, Kuppens, De Boeck, Tuerlinckx, 2013).

The MCMC algorithms for the TRT models and the BCSM were programmed in R. The computation time to complete 10,000 MCMC iterations was less than 5 minutes. This appears to be much faster than the computation times reported by Jiao et al. (2013), who reported a computation time of 6 hours to estimate the one-parameter TRT model using WinBugs. The MCMC estimation method was not computer intensive for the BCSM, and the efficient parameterization of the BCSM makes it possible to fit the model on large-scale tests with large sample sizes.

The BCSM for testlets improves the flexibility of modeling dependences among items within a testlet. This approach can be further explored by considering more complex designs. For instance, additional dependences in the test data can also occur due to different response modes (e.g., different blocks of items require different types of responding), different test domains, or differential item functioning (Paek & Fukuhara, 2015). Future research will focus on developing BCSMs to simultaneously model various types of clustered items, where testlets represent just one way of clustering the items in a test.

References

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

Braeken J, Kuppens P, De Boeck P, Tuerlinckx F. (2013). Contextualized personality questionnaires: A case for Copulas in structural equation models for categorical data. *Multivariate Behavioral Research, 48(6)*, 845-870. DOI: 10.1080/00273171.2013.827965.

Fox, J. P. (2010). *Bayesian item response modeling: Theory and applications*. Springer Science & Business Media.

Fox, J. P., Mulder, J., & Sinharay, S. (2017). Bayes factor covariance testing in item response models. *Psychometrika*, *82*(4), 979–1006.

Hoff, P. D. (2007). Extending the rank likelihood for semiparametric copula estimation. *The Annals of Applied Statistics, 1(1)*, 265-283.

Jiao, H., Lissitz, R. W., & Zhan, P. (2017). A noncompensatory testlet model for calibrating innovative items embedded in multiple contexts. In Jiao, Hong, and Robert W. Lissitz (Eds.), *Technology enhanced innovative assessment: Development, modeling, and scoring from an interdisciplinary perspective.* Charlotte, NC: IAP.

Jiao, H., Wang, S., & He, W. (2013). Estimation methods for one-parameter testlet models. *Journal of Educational Measurement, 50*(2), 186–203.

Klotzke, K., & Fox, J.-P. (2019a). Bayesian Covariance Structure Modelling of responses and process data. *Frontiers in Psychology, 10*, 1675. doi:10.3389/fpsyg.2019.01675

Klotzke, K., & Fox, J.-P. (2019b). Modeling dependence structures for response times in a Bayesian framework. *Psychometrika*, *1-24.* doi:10.1007/s11336-019-09671-8

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement*, *30*(1), 3–21.

Paek, I., & Fukuhara, H. (2015). An investigation of DIF mechanisms in the context of differential testlet effects. *British Journal of Mathematical and Statistical Psychology, 68*(1), 142–157.

Pitt, M., Chan, D., & Kohn, R. (2006). Efficient Bayesian inference for Gaussian copula regression models. *Biometrika, 93(3)*, 537-554.

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News*, *6*(1), 7–11.

Robitzsch, R. (2019). Sirt: Supplementary item response theory models. R package version 3.1-80. Retrieved from: http://CRAN.R-project.org/package=sirt

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.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement*, *26*(3), 247–260.

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. S. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 245–269). Dordrecht, the Netherlands: Kluwer Academic Publishers.

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

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*(3), 185–201.

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: Theory and applications. (ETS Research Report Series, RR 02-02). Princeton, NJ: 2002.

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

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2002). Identification and evaluation of local item dependencies in the Medical College Admissions Test. *Journal of Educational Measurement, 39*(4), 291–309.

Appendix

To implement the MCMC algorithm, samples need to be drawn from the conditional distributions of the parameters θ_i, a_k, b_k, and σ_{γ_d}. The sampling of the model parameters is facilitated by sampling latent responses and by sampling the model parameters given the latent response data.

Consider the BCSM model (Equation [3]), where the augmented data are multivariate normally distributed $\mathbf{Z}_i^d$. The conditional distribution of the augmented data can be simplified by using an explicit form for the covariance matrix $\mathbf{\Sigma}_d$ (Fox, 2010). Let $\mathbf{Z}_{i,-j}$ denote the vector of augmented responses of subject i excluding the j^{th} response. Furthermore, for covariance matrix $\mathbf{\Sigma}_d = \mathbf{I}_d + \sigma_{\gamma_d} \mathbf{J}_d$, let $\mathbf{\Sigma}_{j,-j} = \sigma_{\gamma_d} \mathbf{1}_{n_d-1}^t$ denote the j^{th} row of the covariance matrix excluding the j^{th} value, and let $\mathbf{\Sigma}_{-j,-j} = \mathbf{I}_{n_d-1} + \sigma_{\gamma_d} \mathbf{J}_{n_d-1}$ denote the covariance matrix excluding row j and column j. The conditional distribution of Z_{ij} given $\mathbf{Z}_{i,-j}$ is normal with mean

$$
\begin{aligned}
E\left(Z_{ij}^{d} \mid \boldsymbol{\Sigma}_d, \mathbf{Z}_{i,-j}^{d}, \boldsymbol{a}^{d}, \boldsymbol{b}^{d}, \boldsymbol{\theta}\right) &= \mu_j + \boldsymbol{\Sigma}_{j,-j}\, \boldsymbol{\Sigma}_{-j,-j}^{-1}\left(\mathbf{Z}_{i,-j}^{d} - \boldsymbol{\mu}_{-j}\right) \\
&= \mu_j + \frac{\sigma_{\gamma_d}}{1+(n_d-1)\sigma_{\gamma_d}} \mathbf{1}_{n_d-1}^{t}\left(\mathbf{Z}_{i,-j}^{d} - \boldsymbol{\mu}_{-j}\right),
\end{aligned}
\tag{A-1}
$$

where $\boldsymbol{\mu}_{-j} = a_{-j}^{d}\theta_i - b_{-j}^{d}$ and $\mu_j = a_j^{d}\theta_i - b_j^{d}$, and variance

$$
\begin{aligned}
Var\left(Z_{ij}^{d} \mid \boldsymbol{\Sigma}, \mathbf{Z}_{i,-j}^{d}\right) &= \boldsymbol{\Sigma}_{j,j} - \boldsymbol{\Sigma}_{j,-j}\boldsymbol{\Sigma}_{-j,-j}^{-1}\boldsymbol{\Sigma}_{-j,j} \\
&= \frac{1+n_d\sigma_{\gamma_d}}{1+(n_d-1)\sigma_{\gamma_d}}.
\end{aligned}
\tag{A-2}
$$

For the responses to items in testlet d, the components of variable $\mathbf{Z}_i^d$ are conditionally independent normally distributed, with each mean and variance given in Equations A-1 and A-2, respectively. Each $Z_{id(j)}$ is truncated to be positive if $Y_{id(j)} = 1$ and truncated to be negative if $Y_{id(j)} = 0$.

The parameter σ_{γ_d} is sampled from the shifted inverse gamma distribution with shape parameter $N + g_1$, scale parameter $g_2 + \sum_i \tilde{Z}_{id}^2 / 2$, and shift parameter $1/n_d$, where

$$
\tilde{Z}_{id}^2 = \sum_{k \in d} (Z_{id(k)} - (a_k\theta_i - b_k))^2 / n_d .
$$

The item difficulty parameters can be sampled from a normal distribution given the latent response data. The item difficulty parameters are assumed to be normally distributed with mean μ_b and variance σ_b^2. The conditional distribution of each item parameter k in testlet d is normally distributed with mean

$$
E\left(b_k \mid \mathbf{Z}_{d(k)}, \sigma_{\gamma_d}, \mu_b, \sigma_b^2\right) = \left(\frac{N}{\sigma_{\gamma_d}+1} + \frac{1}{\sigma_b^2}\right)^{-1} \left(\frac{N\left(-\bar{Z}_{d(k)} + \mu_\theta\right)}{\sigma_{\gamma_d}+1} + \frac{\mu_b}{\sigma_b^2}\right)
$$

and variance

$$Var\left(b_k | \; \mathbf{Z}_{d(k)}, \sigma_{\gamma_d}, \mu_b, \sigma_b^2\right) \;=\; \left(\frac{N}{\sigma_{\gamma_d}+1} + \frac{1}{\sigma_b^2}\right)^{-1}.$$

The ability parameter is sampled from a normal distribution. The latent response data $\mathbf{Z}_i$ for test taker i are multivariate normally distributed. The covariance matrix $\boldsymbol{\Sigma}_i$ is a diagonal matrix with D blocks, where block d is equal to $\boldsymbol{\Sigma}_d = \mathbf{I}_{n_d} + \mathbf{J}_{n_d}\sigma_{\gamma_d}$. It follows that the mean and variance of the conditional distribution of θ_i is given, respectively, as mean

$$E\left(\theta_i \mid \mathbf{Z}_i, \mathbf{a}, \mathbf{b}, \mu_\theta, \sigma_\theta, \boldsymbol{\Sigma}_i\right) \;=\; \frac{\mathbf{a}^t \boldsymbol{\Sigma}_i^{-1}\left(\mathbf{Z}_i + \mathbf{b}\right) + \mu_\theta / \sigma_\theta}{\left(\mathbf{a}^t \; \boldsymbol{\Sigma}_i^{-1}\mathbf{a}\right)^{-1} + 1/\sigma_\theta}$$

and variance

$$Var\left(\theta_i \mid \mathbf{Z}_i, \mathbf{a}, \mathbf{b}, \mu_\theta, \sigma_\theta, \boldsymbol{\Sigma}_i\right) \;=\; \left(\left(\mathbf{a}^t \; \boldsymbol{\Sigma}_i^{-1}\mathbf{a}\right)^{-1} + 1/\sigma_\theta\right)^{-1}.$$

The discrimination parameters are also sampled from a normal distribution. The prior distribution of the discrimination parameters is assumed to be normal with mean μ_a and variance σ_a^2. Then, the conditional distribution of each discrimination parameter is normal with mean

$$E\left(a_k | \; \mathbf{Z}_{d(k)}, b_k, \; \boldsymbol{\theta}, \sigma_{\gamma_d}, \mu_a, \sigma_a\right) \;=\; \frac{\boldsymbol{\theta}^t\left(\mathbf{Z}_{d(k)} + b_k\right) + \mu_a / \sigma_a^2}{\dfrac{\boldsymbol{\theta}^t\boldsymbol{\theta}}{(1+\sigma_{\gamma_d})} + 1/\sigma_a^2}$$

and variance

$$Var\left(a_k | \; \boldsymbol{\theta}, \sigma_{\gamma_d}, \sigma_a^2\right) \;=\; \left(\frac{\boldsymbol{\theta}^t\boldsymbol{\theta}}{(1+\sigma_{\gamma_d})} + 1/\sigma_a^2\right)^{-1},$$

respectively.