Item response theory model with testlet effects: A simulation study investigation of effectiveness in small sample sizes

Jeremias Hendrik Marian Wenzel
1st supervisor: Prof. Dr. Ir. Jean-Paul Fox
2nd supervisor: MSc Konrad Klotzke
University of Twente

Abstract

In this article, two different item response theory (IRT) models with testlet effects are compared. Testlets structures account for item interrcorrelations in sets of items, so that conditional independence for a test can be assumed, even if groups of related items are used in the testconstruction. The conditional testlet model adds a testlet parameter to the basic IRT model, that models the dependencies among response observations to the same testlet items. The marginal testlet model, by contrast, models them directly as a covariance parameter in the covariance matrix of the distribution of the error. This leads to better parameter estimation for small dependencies. The precision of the parameter estimates under the two models are evaluated for data simulated under the conditional testlet model. The simulation study uses different conditions, varying in number of participants, number of items included per testlet and amount of variance in the testlet effects. It will be shown that under the marginal testlet model, parameter estimates can be obtained with the same precision as under the conditional testlet model for large sample sizes. Furthermore, under the marginal testlet model, better parameter estimates are obtained for conditions with small testlet variances than under the conditional testlet model.

Item response theory model with testlet effects: A simulation study investigation of effectiveness in small sample sizes

Introduction

Item response theory is a form of test scoring, which includes the latent ability of a given examinee and the difficulty of the item, in order to estimate the chance of success on a test item. In contrast to classical test theory, which assumes all items to have equal difficulty, the difficulty is estimated per item. This makes it possible to administer test questions in random order. The basic item response theory (IRT) assumes all items to have conditional independence (CI). This means that no items are related to each other given the value of the latent variable. This becomes problematic in the case of so called testlets. Testlets are groups of questions that are inherently related to each other. A common example are multiple questions in a reading comprehension questionnaire. Because it takes some time to read a long section of text, it is useful to ask multiple questions about it. Therefore, all items are pertaining to a single piece of text. It makes intuitive sense that these questions are not independent from each other. Answering correctly on one of them is inherently related to the answering of the others, thus leading to a possible wrong evaluation of the examinees proficiency. Under the basic IRT model, an examinee who answered three questions correctly from one reading comprehension excerpt would be evaluated equally to a different examinee who answered three individual questions correctly. Their proficiency would be estimated to be the same (given equal difficulty for the items), even though they each showed their knowledge of only one, compared to three pieces of information, respectively. The term testlet was first introduced by Wainer and Kiely (1987), as a way of adressing this problem.

Conditional testlet model

To address the issue of sets of interrelated items, Bradlow, Wainer and Wang (2007) have developed the testlet model. This model will hereafter be reffered to as the conditional testlet model. This model is an extension of the basic item response theory model, with the inclusion of a testlet effect parameter. This effect represents the correction of a person's proficiency, due to the items inter-correlation in response observations. The testlet parameter adds an additional contribution on the success of a correct response. This additional contribution is added with a

negative sign, thus reducing the chance of a correct response of the participant, if the item in question is nested within a testlet. The equation represents the chance that a person i scores correct on a test item j, indicated as $Y_{ij} = 1$:

$$P(Y_{ij} = 1) = \Phi(a_j \cdot \theta_i - \gamma_{id(j)} - b_j), \tag{1}$$

where Φ is the cumulative normal distribution function, a_j is the discrimination parameter, θ_i is the ability parameter of a person, and b_j the difficulty parameter per test item. These parameters are equivalent to the parameters of the 2-PL item response theory model. The parameter $\gamma_{id(j)}$ represents the added testlet effect, representing the dependency among responses. The indices $id(j)$ represents person i item j, nested in testlet d. For the responses of person i on items j and j', that are within the same testlet $d(j) = d(j')$, $\gamma_{id(j)}$ would be the same for that person. This shared testlet effect for items within a testlet gives rise to a correlation between item responses of a person for that testlet. The distribution of $\gamma_{id(j)}$ is assumed to be normal with a variance, which is set a priori to analyzing the data, $\gamma_{id(j)} \sim N(0, \sigma^2_\gamma)$. The testlet effects are thus modeled as the variance of this parameter. For σ^2_γ, the testlet effect variance, a noninformative inverse gamma prior was specified with parameters set to 0.01. For further details on the testlet model, see Wainer, Bradlow and Wang (2007)

Given the inclusion of the testlet factor, conditional independence is assumed to hold again. The resulting estimation of the parameters of the model are thus assumed to have only random error and no unexplained error due to the interdependence of groups of items. There are four possible shortcomings to this model: First, because a testlet distribution has to be chosen before the analysis of the data, it has an influence on the actual estimated testlet effect. Therefore, the choice of this prior distribution has an impact on the eventual effect on the success probability of an individual. This does not make an impact for large numbers of participants (e.g.: $N > 1000$) and large numbers of items per testlet (e.g.: 10-15), but makes the method unreliable for smaller groups of participants (e.g.: $N < 100$) and small numbers of items per testlet (e.g.: < 10). Both of these factors limit the amount of information about the dependencies among response observations to estimate the testlet effects. Second, the chosen standard deviation of the testlet effect cannot be zero (then there would be no testlet effect at all) or negative (a variance cannot be negative). Therefore, it must be decided beforehand, whether a testlet effect is present. Once the

normal distribution is specified, a positive testlet effect is assumed. Thus, the testlet model cannot be used to examine whether the supposed interrelation of items in a testlet are actually supported by the data. Another possible shortcoming results from this. The testlet variance cannot be estimated to be negative. As a result, very small testlet effects might be systematically overestimated. The variance of the testlet effect is restricted to be greater than zero, and due to this lowerbound, the distribution of the testlet variance is highly skewed to the right. As a result, the variance of the testlet effects is easily overestimated leading to more variability in testlet effects than supported by the data. Ultimately, this would mean that correct answers on testlet items would be unjustly devalued in their prediction of the chance of success. Lastly, using this model, one testlet variance is defined for all testlets. The underlying assumption is that all testlets have items that are equally related to each other. This might seem like an assumption that is difficult to adhere to in real-world applications.

Marginal testlet model

Due to these shortcomings, an alternative testlet model is investigated to deal with testlet effects, hereafter referred to as the marginal testlet model. In the marginal testlet model, the dependencies among response observations to the same testlet items are not modeled with a testlet parameter. They are directly modeled as a covariance parameter in the covariance matrix of the distribution of the errors. This is accomplished using the latent response formulation for the two-parameter IRT model under a probit link function. A multivariate normal distribution is assumed for the latent responses with as mean term $a_j\theta_i - b_j$, and as covariance matrix a common covariance among item responses assigned to the same testlet. The common covariance parameter resembles the dependency among responses modeled by a testlet parameter in the mean term in the conditional testlet model. In the conditional testlet model, the covariance between item responses in the same testlet is represented by the variance of the testlet effect. In the marginal testlet model, the dependencies can also be modeled by a common covariance parameter, but this parameter does not need to be restricted to be positive, which is a natural restriction on a variance parameter. As a result, the covariance among responses to the testlet items are less restricted than the implied dependency by the testlet variance parameter. It is to be expected that the marginal testlet model will perform better when the covariance is small and close to zero. The lowerbound

of zero of the testlet variance parameter in the conditional testlet model can easily lead to an overestimation of the testlet variance when the covariance is close to zero. The lowerbound will also lead to a skewed posterior distribution of the variance parameter, which can lead to estimation problems. In the marginal model this lowerbound restriction of zero does not apply, which leads to a more symmetric posterior distribution of the covariance parameter. This also facilitates less issues in estimating the marginal testlet parameters. In the marginal model the testlet effect parameters are not included in the mean term and can also not be directly estimated. When the testlet effects are of specific interest, a post-hoc estimate can be obtained by computing the testlet effect from the fitted residuals given the estimated marginal model parameters.

Motivation research

The need for reliable testing methods has never been bigger. Tests are getting increasingly modular (Wainer, Bradlow, and Wang, 2007). Methods like computerized adaptive testing (CAT) are gaining more popularity. In these methods, different examinees are not necessarily presented the same items in the same test, in order to estimate their proficiency. Therefore, issues of item local dependence are a pressing issue, that need to be addressed. Testlet structures give the opportunity to easily incorporate convenient and efficient test designs, without the worry of violating conditional independence assumptions. It makes sense to ask multiple questions pertaining to one topic or excerpt from both a perspective of efficiency and ease of use for the test-designer. With a testlet model, these kinds of test-designs can be easily incorporated and used. The current, conditional testlet model is limited to large scale testing with thousands of participants. This limitation makes it only useful for big tests, like the Standardized Aptitude Test (SAT) in the US. A testlet model that works reliably under conditions with little participants could make it useful for many smaller-scale applications. This study aims to investigate the usefulness and relative advantage of the marginal testlet model to be applied in this role. Furthermore, the goal is to provide advice about the areas where additional research into the marginal testlet model is useful. In this simulation study, the marginal testlet model is compared to the conditional testlet model. Using a simulation study affords the advantage that the true parameter values are known. In this way, accuracy of the estimation method is measurable, something that is not possible in an applied setting, where the true values are always unknown. The sampled data for the simulation is

based on the conditional two-parameter testlet model. The main interest of this study lies in the comparison of the conditional and marginal testlet model. First, it is investigated whether the marginal model performs as well as the conditional testlet model. Therefore, parts of the simulation will be replicating the conditions used by Bradlow and Wainer in their simulation study (Wang, Bradlow and Wainer, 2002). Both models are evaluated, based on how accurate the parameter estimates obtained under the models are, when compared to the true values of the simulated data. Furthermore, because of the possible shortcomings of the conditional testlet model, as described above, estimates obtained under both models will be compared in conditions with small testlet variances. It is expected that the marginal testlet model will perform better under these circumstances. Additionally, the estimation methods will be run with less available information on the underlying data structure. This means that a smaller number of participants and less items per testlet will be used. The expectation is that under the marginal testlet model, better results are delivered than under the conditional testlet model.

Method

Data generation

A population distribution is defined, modeled under the conditional testlet model. Consequently, values are drawn from this distribution at random, forming the datasets of response observations. These datasets mimic the response observations that would be gathered in an applied setting, if we were to assume that the conditional testlet model structure was true for a real population. In the case of this simulation, this consists of a matrix of binary responses, since only dichotomous items were used. An observation of 1 indicates success and an observation of 0 no success of a participant on a test-item. This is repeated for each condition and replication. The datasets are then analyzed under the assumptions of the two models.

The following population distributions were asserted for the generation of the data sets:

$$\begin{aligned} a_j &\sim N(0,1) \\ b_j &\sim N(0,1) \\ \theta_i &\sim N(0,1) \\ \gamma_{id(j)} &\sim N(0,\sigma_\gamma^2), \end{aligned}$$

where $N(\mu, \sigma^2)$ donates a Normal distribution with mean μ and variance σ^2. The variance of the testlet parameter $\gamma_{id(j)}$ is adjusted based on the different conditions. This 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 as big as the ability variance in that participant.

Computation

The simulation study was executed using the statistical program R. Both the data generation and parameter estimation under the conditional and marginal testlet model were performed using custom written R-Scripts. A seed was set at the start of each conditions run, insuring replicability of the results. Each estimation started with the generation of a dataset of response observations, which was then used by, first, the estimation method under the conditional testlet model and then the marginal testlet model to obtain the posterior parameter estimates of interest.

MCMC

A Markov chain Monte Carlo (MCMC) procedure was used. This procedure samples values from the posterior distribution of specified model parameters (Wainer, Bradlow and Wang, 2007). Using the simulated response observations and the prior distributions, the MCMC generates iterations of parameter estimates of the model parameters (e.g. the difficulty parameter, denoted earlier as b_j), until it converges to the stationary distribution. This means that the algorithm samples values from the target posterior distribution of the model parameter. When convergence is reached, additional samples from the posterior distribution are drawn, which are used for inference. This has the advantage that further inferences, like computing the mean of the posterior distribution of a parameter, can be easily made (Wainer, Bradlow and Wang, 2007). According to Wainer, Bradlow and Wang (2007), convergence for the 2-PL conditional testlet model is usually reached within 10,000 iterations. MCMC sampling consists of two phases. First, the burn-in fase. In this fase, the iterations approach the stationary distribution from the set starting point. This fase is not used for the model estimation and is discarded. The second part are the iterations used for inference. The chosen ratio between the two phases is a trade-off between computation time and Monte Carlo simulation error (Wainer, Bradlow and Wang, 2007). In this study, the first 1,000 iterations were discarded as the burn-in fase, leaving iterations 1,000 to 10,000 for the estimation

of the parameters.

Preliminary testing

In order to determine the conditions that could be used in the study, preliminary tests were performed. The goal was to find the lower limits at which the algorithm under the conditional testlet model still produces reliable parameter estimates. For this purpose, different conditions with small sample sizes were observed. The Heidelberger and Welch's convergence diagnostic (Plummer, Best, Cowles and Vines, 2006) was used to evaluate the algorithms performance under certain conditions. The estimation method under the conditional testlet model showed more convergence issues than under the marginal one during the initial tests. Mainly, number of participants, testlet variance and number of items per testlet influenced the functioning of the MCMC algorithm, leading to them being chosen for the simulation study. It should be noted that these preliminary tests were conducted using the MCMC procedure for 1000 to 2000 iterations only, due to the unfeasibility of running it for long periods of time. Ultimately, these were used to gain an impression of the functioning of the two algorithms under different conditions, and were used to inform the choice of conditions of the simulation study.

Design

This simulation study was performed for two reasons. First, to show that the estimation method under both the marginal testlet model and the conditional testlet model function equally well for conditions with large numbers of participants. Secondly, the goal was to investigate whether the marginal testlet model performs better in situations with little information available. The simulation values were defined similarly to the ones used by Wang, Bradlow and Wainer (2002). Each condition was replicated 50 times. The number of items was set to 30 for each condition. Three factors were chosen for manipulation. First, the number of participants was set to 1,000, 500, and 200. For conditions with 200 participants, accurate parameter estimates were still obtained under the conditional testlet model in the preliminary testing. Second, the number of items per testlet was manipulated. The choice was made for 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 an applied setting,

and were the focus of this current study. Therefore, the choice was made to use a variance of .1, .05, and .01 for the testlet effect. The resulting full-factorial design consisted of $3 * 3 * 3 = 27$ conditions. Because this would have been too computationally time-consuming, the choice was made to use a Latin-square design. Instead of varying the third factor over all possible combinations of the first two factors, it was used only once for every factor. The resulting simulation design had 9 distinct conditions. In table 1, the whole design is shown. The numbers 1 to 9 in parenthesis refer to the numbering of the nine conditions.

Table 1

Table of Simulation Design

Variance of the Testlet Effect		# items per testlet		
		5	10	15
# participants	1000	.05 (1)	.1 (2)	.01 (3)
	500	.1 (4)	.01 (5)	.05 (6)
	200	.01 (7a & b)	.05 (8)	.1 (9)

Note. The latin square design limits the number of conditions needed, but still maximize the number of combination of conditions observed. The numbers in brackets () indicate the number of the condition. Condition 7 was run two times, with different prior specifications for the testlet parameter under the conditional testlet model. They are indicated by 7a and 7b.

In order to investigate the effect of different priors on the parameter estimation under the conditional testlet model for conditions with little information, condition 7 was ran a second time, with different prior specifications. Condition 7a was run with the noninformative prior for the testlet parameter variance set to 0.01. Condition 7b, on the other hand, was run with an informative prior specification for the algorithm of the conditional testlet model. This prior for the testlet parameter variance was set to 1. All other parameters, like number of participants, true testlet variance, and number of items per testlet remained the same.

Criteria

Through sampling from the posterior distributions of model parameters, parameter estimates were obtained. Calculating bias and mean squared error (MSE) of these parameters shows the accuracy of these parameter estimates. It is an indication of how well the two models each describe the data. The bias and mean squared error (MSE) were calculated for different model parameters obtained under the two models. Where applicable, the results were averaged across the test items and participants, to make the comparison more straightforward. Bias was calculated as the average difference between the true and estimated value;

$$Bias(\hat{b}_j) = \frac{1}{50}\sum_{r=1}^{50}(\hat{b}_{jr} - b_{jr}), \tag{2}$$

where $\hat{b}_{jr}$ is the estimated difficulty parameter for item j in replication r. b_j represents the true value of the difficulty parameter for item j in replication r. MSE was calculated similarly, with:

$$MSE(\hat{b}_j) = \frac{1}{50}\sum_{r=1}^{50}(\hat{b}_{jr} - b_{jr})^2. \tag{3}$$

Bias and MSE were calculated for the discrimination parameter (a), the difficulty parameter (b), the ability parameter (θ). Furthermore, bias and MSE was calculated both for the variance of the testlet effect and the estimated individual testlet effects. Additionally, the mean absolute deviation of the individual testlet effects was calculated. This was done by:

$$mean\,absolute\,deviation(\hat{\gamma}) = \frac{1}{50}\sum_{r=1}^{50}(\,|\hat{\gamma}_r - \gamma_r|\,). \tag{4}$$

Mean absolute deviation averages the absolute differences, regardless of their sign. This way, the overall distance from the true values were estimated. It was expected that the estimates obtained under the conditional testlet model would have lower overall deviation from the mean. It was of interest to investigate this. The estimated testlet effects obtained under the conditional testlet model were rescaled to make the comparison of MSE, bias, and mean absolute deviation between the two models fair. The conditional testlet model uses regularized estimation. The estimated testlet effects are a combination of both the data information and the smoothing influence of the prior distribution. The marginal model, on the other hand, uses unregularized estimation of the testlet effects. No smoothing via a prior distribution is needed. Because of this difference, the

MSE and bias of the two models are not on the same scale, artificially lowering the MSE and bias under the conditional testlet model by comparison. In order to make a fair comparison possible, the estimated testlet effects under both models were rescaled to be on the same scale. Outcomes on the testlet estimates given below are based on the rescaled values.

Another indicator of how well the models describe the data is the coverage rate. A Highest Postrior Density (HPD) interval was used. The interval gives an indication of the amount of times that the true value lies within the predicted interval. Both models are designed to have a 95% coverage rate. If this is lower, the true values do not lie within the predicted interval enough times. The lower the coverage rate, the less often we would find the true value to lie within the 95% HPD, given many replications. If the coverage is low, the model would claim to find the underlying true value more frequently than is actually warranted. A low coverage rate indicates a problem with the models functioning. It is predicted that the coverage rate under the marginal testlet model will be higher than the true 95%. This is not indicative of bad model functioning, since under the marginal testlet model observations are not restricted to be positively correlated within a testlet. This leads to a wider confidence interval. As the data is generated under the conditional testlet model, the coverage does not match up exactly.

The estimation of the MSE and bias was based on mean estimation. It was expected that the conditional testlet model would have a skewed posterior distribution of the variance parameter. If this is the case, mean estimation could lead to inaccurate estimations, as the mean would not represent the center of the distribution. Therefore, results were also obtained when using a mode estimation method. If the differences between the two estimations are large, this would indicate that the posterior distributions are skewed and the results may be imprecise. This was predicted for the conditional testlet model under small testlet variances. The Asselin de Beauville mode estimator (Poncet, 2012) was used to estimate the mode, and results obtained using both methods were compared.

Results

Not all of the conditions worked as intended. In condition 6 and 9, under the marginal testlet model, the estimation of the discrimination parameter led to problems. This was due to some very high covariances that occurred, and posed problems for the estimation method. These

Table 2

Estimated Testlet Variance for the Conditional and Marginal Model

#condition	true variance	estimated variance	
		conditional	marginal
1	.05	.0445	.0539
2	.1	.0982	.1007
3	.01	.0098	.0112
4	.1	.0953	.1052
5	.01	.0076	.011
6	.05	.0496	.0496
7a	.01	.0027	.0089
7b	.01	.0609	.0108
8	.05	.0332	.0527
9	.01	.0938	.0985

Note. The average estimated testlet variances were recovered from the conditional and marginal testlet model. A value closer to the true variance indicates better model performance

conditions were run again with the discrimination parameter fixed to 1. We think that this does not influence the outcomes in any meaningful way, though.

In all ten conditions parameter estimates were obtained. For the estimation of the model parameters a, b, and θ, very little to no differences were found between the estimation methods under the two models. Estimates for bias and MSE of these parameters are displayed in the Appendix, Table 8 and Table 7 respectively. Small differences were observed between the bias of a, but we think they are not big enough to be relevant.

The estimated testlet variance per condition is shown in Table 2. We see that under the conditional testlet model, the true testlet variance was generally underestimated, except for condition 7b, which will be further explored later on. Especially in conditions 7a and 8 very low estimates were obtained. For condition 7a this estimated testlet variance was .0027, with the true testlet variance at .01. In condition 8, the estimate was .0332 with a true value of .05. Under the

Table 3

MSE of True Parameters and Estimated Posterior Mean for the testlet variance and Testlet Effects

		testlet variance		estimated testlet effect	
con #	parameters:	conditional	marginal	conditional	marginal
1	1000, .05, 5	.0003	.0001	.0708	.0708
2	1000, .1, 10	.0001	.0001	.1079	.108
3	1000, .01, 15	.0001	.0001	.0162	.0163
4	500, .1, 5	.0005	.0004	.1248	.1247
5	500, .01, 10	.0001	.0001	.0165	.0167
6	500, .05, 15	.0002	.0002	.063	.063
7	200, .01, 5	.0001	.0004	.017	.0171
8	200, .05, 10	.0006	.0003	.065	.065
9	200, .1, 15	.0008	.0007	.1055	.1052

Note. The MSE for the testlet variance and the estimated testlet effect, estimated under the conditional and the marginal testlet model.

marginal testlet model, on the other hand, the testlet variance tended to be slightly overestimated. Only in conditions 6, 7a, and 9, the testlet variance was underestimated. In general, the estimates were very precise under the marginal testlet model.

We see that the parameter estimates, under the marginal testlet model, are slightly different for the conditions 7a and 7b (Table 2, Table 5). As the estimates were obtained while using a constant seed, we would expect the exact same data for both conditions under the marginal testlet model, as the only difference between the two conditions (informative vs. uninformative prior for the testlet effect) does not apply to the marginal testlet model. But because the generated datasets were analysed under both testlet models within the same run of the program, the changes in estimation under the conditional testlet model have an impact on the results of estimation under the marginal testlet model. Because the estimations rely on random draws, these slight variations in results are expected.

The differences in MSE of the testlet variance was very small, as can be seen in Table 3. As we can see in Table 4, the bias of the testlet variance is very small under both models. Under the

Table 4

Bias of True Parameters and Estimated Posterior Mean for the Testlet Variance and Testlet Effects

		testlet variance		estimated testlet effect	
con #	parameters	conditional	marginal	conditional	marginal
2	1000, .1, 10	-.0018	.0007	-.0015	-.0015
3	1000, .01, 15	-.0002	.0012	.0002	.0003
4	500, .1, 5	-.0047	.0052	-.0009	-.0008
5	500, .01, 10	-.0024	.001	-.0004	-.0002
6	500, .05, 15	-.0004	-.0004	-.0008	-.0007
7	200, .01, 5	-.0073	-.0011	.0002	.0003
8	200, .05, 10	-.0168	.0027	.0028	.0026
9	200, .1, 15	-.0062	-.0015	-.0013	-.0013

Note. The bias for the testlet variance and the estimated testlet effect, estimated under the conditional and the marginal testlet model.

conditional testlet model, bias was smaller than under the marginal testlet model for all conditions. In fact, the bias was negative, ranging from $-.0168$ in condition 8 to $-.0002$ in condition 3. For the testlet effects themselves, we see lower values of bias under the conditional testlet model in conditions 4, 5, 6, and 7. This means that under the conditional testlet model, the estimates of the testlet effects and the testlet variance were slightly closer to the true values. For the testlet effects, the mean absolute deviation is shown in Table 5. There are very little differences. The estimates obtained under the marginal testlet model have nearly the same testlet estimates.

The coverage rate under the conditional testlet model was not satisfactory for most conditions. As can be seen in Table 6, under the conditional testlet model, a lower coverage rate for most conditions was found. Only in condition 2 was the coverage rate higher than the target 95%, with a coverage of 100%. The next highest is in condition 3, with a coverage of 90%. In the other conditions, low coverage rates were found. This indicates that the true value did not fall within the estimated confidence interval.

As predicted, under the marginal model, the coverage rate that was too high. For all

Table 5

Mean Absolute Deviation of the Estimated Testlet Effects

#condition	conditional	marginal
1	.212	.212
2	.2623	.2624
3	.1014	.1019
4	.2825	.2822
5	.1027	.1031
6	.2008	.2007
7	.1041	.1041
7b	.1051	.1052
8	.2039	.2038
9	.2597	.2589

*Note.*The mean absolute deviation. The rescaled effects showed nearly no difference between estimates under the two models.

conditions but 6 and 9, the coverage rate was 100%. In condition 6 and 9, the coverage rate is 94% and 96% respectively. These were the conditions where the discrimination parameter was adjusted, which could explain the difference to coverage rates in the other conditions, under the marginal testlet model.

In most conditions, the algorithm under both models was able to perform well and estimate accurately. To illustrate this, Figure 1 is presented. The plot shows the distribution of the posterior sampling estimates of the testlet variance of one replication in one condition. Figure 1 displays the 14th replication of condition 2. This plot does not reflect the overall mean of the estimations, but is used to visualize the results. The replication was chosen as representative of the general trend. The higher the density at a particular value of the testlet variance, the more estimates were obtained by the MCMC method for that particular value. The solid line shows the density of the estimated testlet variances under the conditional testlet model. The distribution has a peak at

Table 6

Coverage of the 95% Confidence Interval

#condition	conditional	marginal
1	.78	1.00
2	1.00	1.00
3	.90	1.00
4	.86	1.00
5	.76	1.00
6	.88	.94
7a	.10	1.00
7b	0	1.00
8	.76	1.00
9	.86	.96

Note. Convergence was calculated using a HPD function. The coverage under the conditional testlet model is too low. Especially condition 7a and 7b. Under the marginal model, coverage is higher than the target coverage.

roughly .1, which is the true value of the testlet variance in condition 2. The distribution has a symmetrical shape and a good amount of variance. The distribution of posterior samples under the marginal testlet model is indicated by the dashed line (Figure 1). It also shows a symmetrical distribution. The sampled values are spread out more along the x-axis, and the density is lower than that estimated under the conditional testlet model. The higher and narrower the distribution, the more certainty there is in the model about the position of the true value. Under the marginal testlet model, the spread of estimated values is bigger, indicating less certainty. For the algorithm to obtain valid posterior estimates, it must sample from the target posterior distribution, meaning it must sample values around the mean with a variance. Figure 2 is another way of visualizing

this. It displays the draws of the estimation algorithm from the posterior distribution under the two models. The MCMC chain runs from the left to the right estimating values. For the algorithm to obtain good parameter estimates, the MCMC chain must sample values around the true value of the parameter in question, with a variance (along the y-axis in Figure 2). The estimated values for the algorithm under the marginal testlet model show this very well. We see that the mean of the sampled values lies at around .1 and observations are sampled with a variance. Estimated values reach form around .05 to .15. Also, no clear trends away from these values can be seen over the course of the draws, indicating convergence. The algorithm under the conditional testlet model also performs well. We can see how it samples values with less variance (the spread of the values along the y-axis is smaller). The mean of the sampled values lies around .1 and there are no drastic spikes.

.

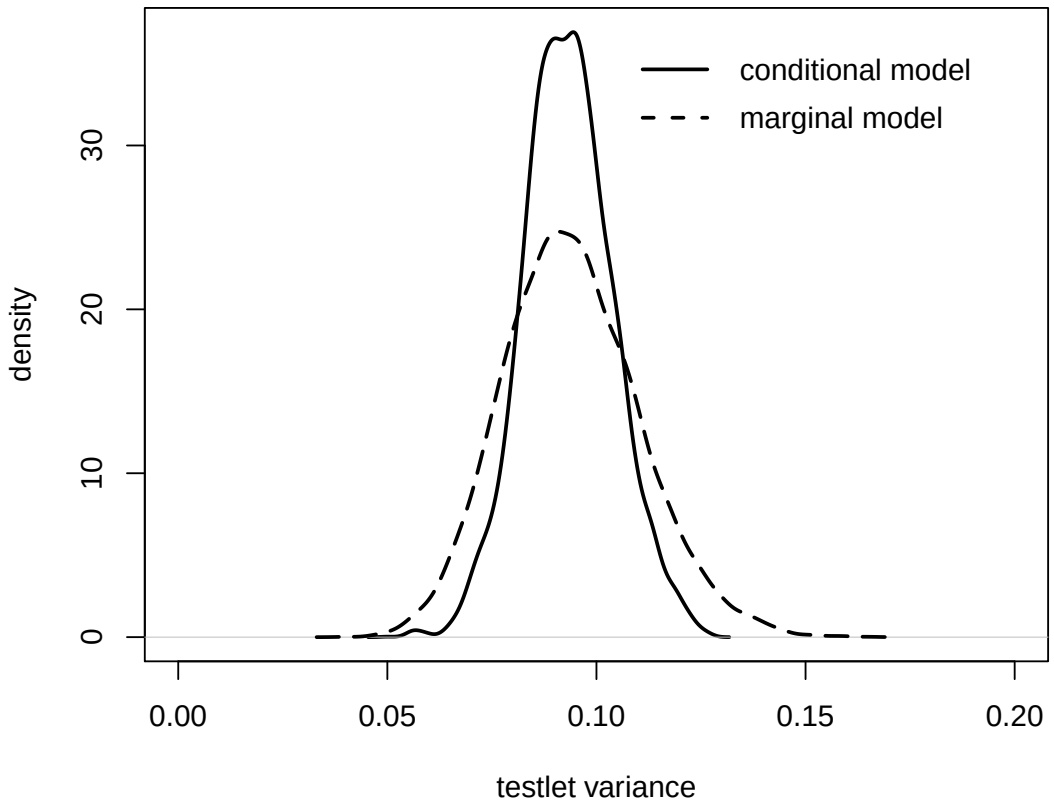

Figure 1. Density of the sampled testlet variance estimates from the posterior distribution under the conditional and marginal testlet model. Under the conditional testlet model, estimates show a higher density and a narrower variance than under the marginal testlet model. The 14th replication of condition 2 was used.

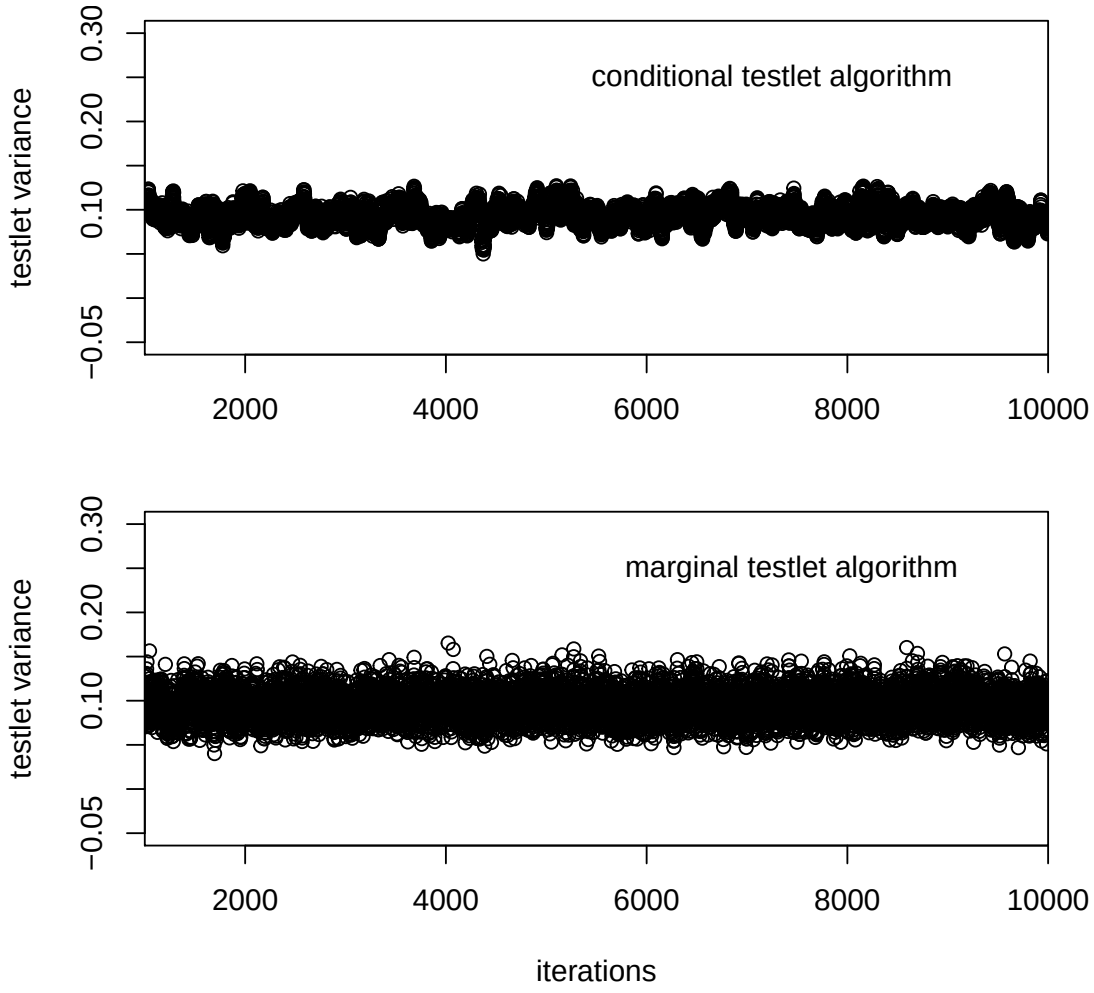

Figure 2. The Figure shows the samples from the posterior distribution of the testlet variance. The MCMC chain goes from left to right. The x-axis goes from 1000-10,000, as the first 1000 iterations are used to reach the posterior distribution. The 14th replication of condition 2 was used.

For some conditions, we could not obtain reliable estimates for the conditional testlet model. This varies per condition, with some more affected than others.

In condition 7a, we were not able to obtain reliable estimates under the conditional testlet model. This is indicated by the low estimated testlet variance (Table 2) and the extremely low coverage rate (Table 6). Condition 7 had a low amount of information with only 200 participants, 5 items per testlet (Table 1). Furthermore, the true testlet variance was set to 0.01, which means that there is less information about the covariances in the data. To illustrate the issue, we look at

Figure 3. Again, replication 14 is chosen, now of condition 7. The dotted line represents the distribution of estimated testlet variance values, under the conditional testlet model, in condition 7a with the noninformative prior specification. The peak of the distribution is far above the bounds of Figure 3, at a density above 300. Due to the lowerbound of zero of the estimates that can be drawn under the conditional testlet model, there is a sharp cut-off of the curve at a testlet variance of zero. The distribution is not symmetrical and shows very limited variance in estimated values. All of the sampled values lie just above zero. The solid line in Figure 3 represents the distribution of estimated values from the posterior distribution under the conditional testlet model, specified with an informative prior. The distribution is symmetrical and shows good variance in the estimated values, similar to the curve observed in Figure 1. Yet, the mean of the distribution lies at around .06. This is a lot higher than the true testlet variance of .01, in the simulated dataset. The parameter estimate is thus overestimating the true value. This is also reflected in Table 2 for all replications combined. The mean estimated testlet variance of condition 7b is .061. With both, a non-informative prior specification, and an informative prior specification, we could not obtain good estimates of the testlet variance under the conditional testlet model. Figure 4 shows the traceplot for the estimates sampled from the posterior distribution of the testlet variance. The top part of the figure shows the values drawn in the MCMC chain under the conditional testlet model in condition 7a. The line runs almost straight, with barely any distribution of observations. This leads to the problems we have seen before. In the middle part of the figure, condition 7b is shown, under the conditional testlet model. Here, the distribution of estimates is similar to that in Figure 2. As we could see in Figure 3, the mean of the distribution of testlet variances lies too high, above .05.

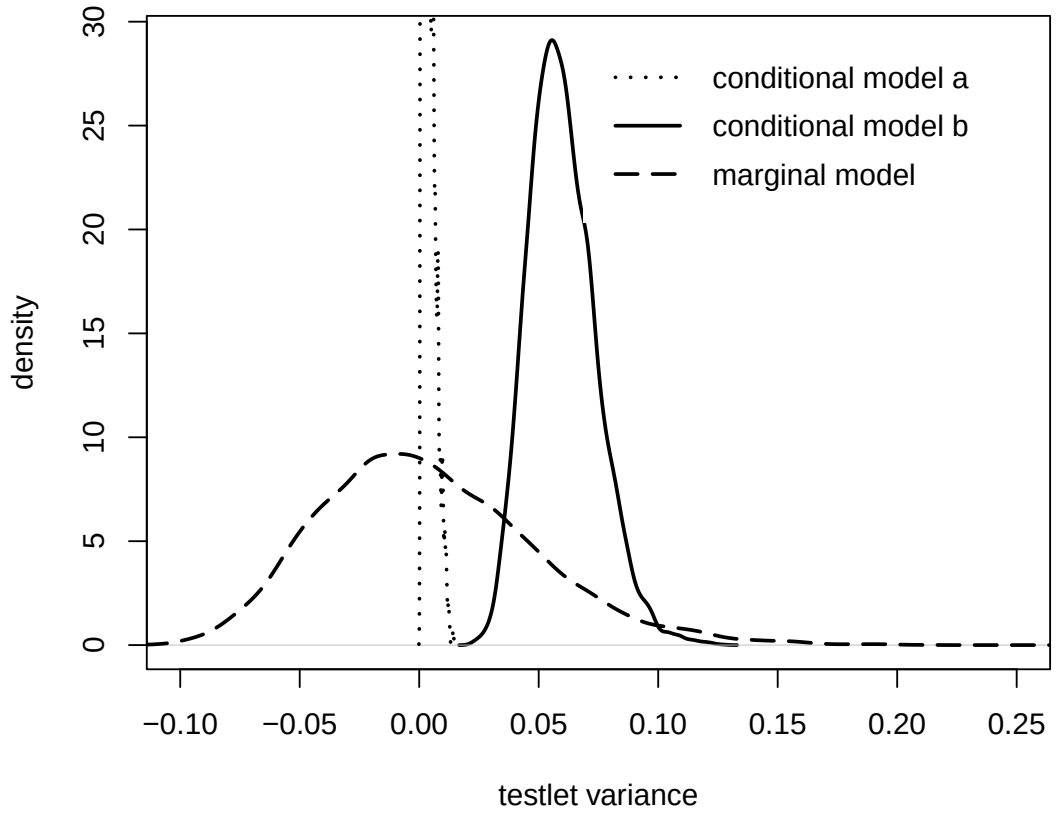

Figure 3. Density of the sampled testlet variance estimates from the posterior distribution under the conditional and marginal testlet model. The 14th replication of condition 7 was used. The peak density of the distribution under the conditional testlet model for condition 7a lies outside the scale of this figure, above 300.

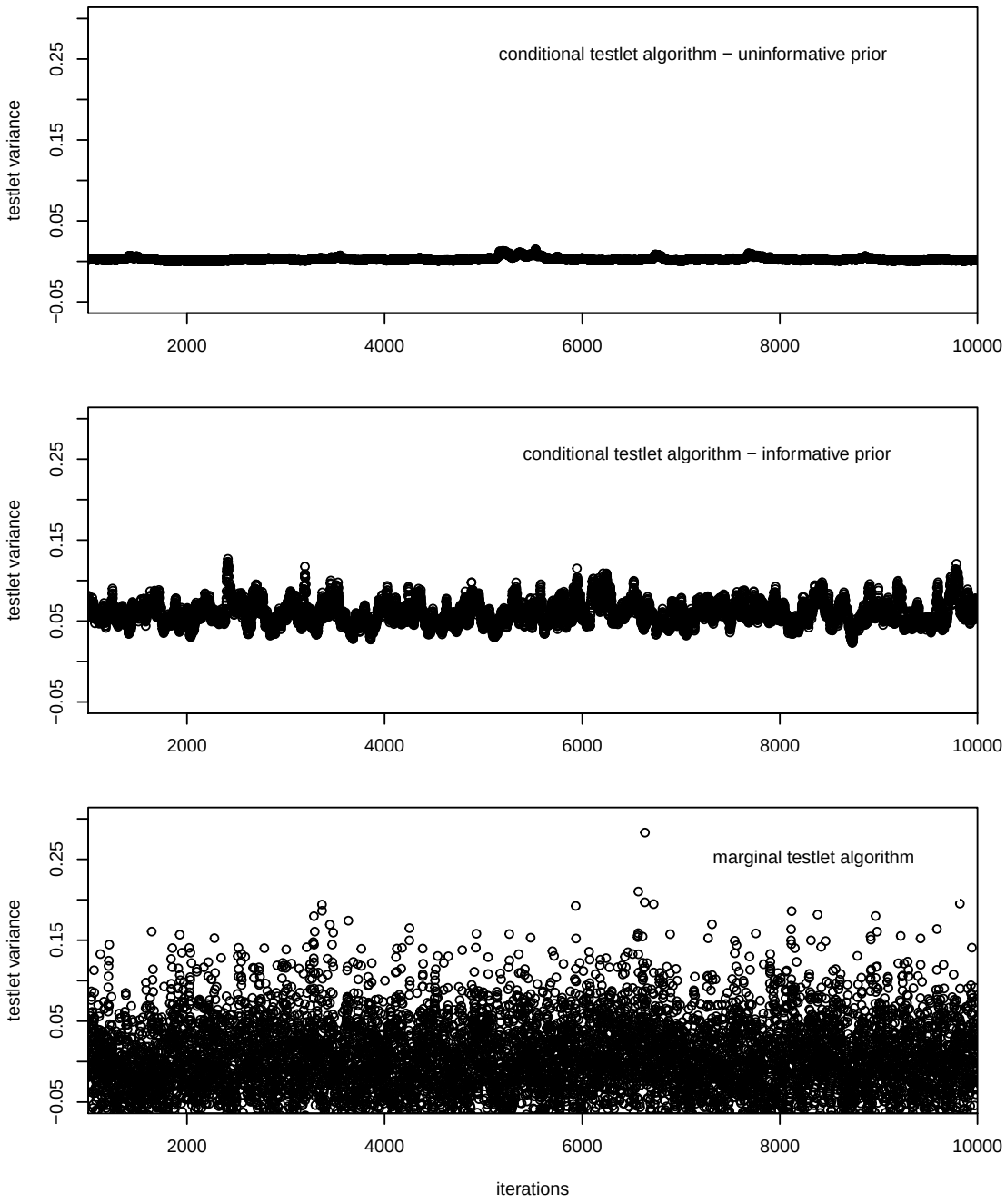

Figure 4. The Figure shows the samples from the posterior distribution of the testlet variance. It is shown at the top how, with the specification of a non-informative prior, the algorithm cannot sample from the posterior distribution. The 14th replication of condition 7 was used.

Under the marginal testlet model, on the other hand, reliable estimates of the testlet variance were obtained. In Figure 3, the dashed line illustrates the distribution of estimated testlet variances under the marginal testlet model. Values below and above zero were drawn. The curve

is symmetrical, with its mean around zero. The highest density is only just below 10 and the distribution is larger. Compared to the distribution of values in Figure 1, under the marginal testlet model, the distribution is much wider. This is not surprising, as there is much less information available, which introduces more uncertainty into the estimation. Again, the same can be seen in the density plot in Figure 4 in the bottom figure. The algorithm draws estimates from the posterior distribution of the testlet variance with a wider distribution. We can see how under the marginal testlet model, estimates can be sampled lower than zero.

In all of the replications of condition 7a, the algorithm failed to obtain good parameter estimates for the testlet variance. In other conditions, like condition 8, only in some of the replications similar difficulties were observed. Some of the replications performed well, similar to the distributions shown in Figure 1, and in others the distribution failed to show good variance.

The mean and mode estimations for the MSE and biases showed very small differences. This means that the mean and the mode did not vary from each other strongly. This did not support the hypothesis that the conditional testlet model would show strongly skewed posterior distributions. Therefore, it was viable to use mean estimation for the bias and mean squared error.

Discussion

This article presents evidence that, first, under the marginal testlet model, accurate parameters for data generated under the conditional testlet model can be estimated, and, second, that the marginal testlet model functions better than the conditional testlet model under conditions with little information and small testlet variances.

The first main point is that for all of the conditions, the same level of accuracy under the marginal testlet model as under the conditional testlet model in estimating the model parameters was reached. The marginal testlet model achieved this without estimating an extra testlet effect parameter. This is an advantage over the conditional testlet model, as it decreases the amount of parameters that have to be estimated. This is relevant for real-world applications with large computations, where the number of parameters quickly accelerates. Thus, a leaner model is preferable. Individual testlet effect sizes are of practical importance, as they permit the prediction of individual responses. It is shown that under the marginal testlet model, these are estimated with the same accuracy as under the conditional testlet model, by using a post-hoc approach.

The second main point of this paper was to show the marginal testlet models applicability for datasets with less available information. It was shown that under the conditional testlet model, a point is reached at which its outcomes are highly reliant on the testlet effect prior specification. Condition 7a and 7b showed how we are not able to obtain reliable parameter estimates if the information in the dataset is too little. It was shown that under the marginal testlet model, reliable estimates are delivered in conditions with few participants, small testlet variance, and few items per testlet. The conditional testlet model did function better in some regards than was expected beforehand. The issues in obtaining reliable posterior estimates under the conditional testlet model did not have a clearly visible effect on the bias and MSE of model parameters. Nevertheless, the coverage rate and the estimated testlet variances, as well as the density plots, are evidence of the limited usefulness of the conditional testlet model.

An issue with the current research was the fact that the marginal testlet model had problems functioning under rare conditions. It is our belief, though, that this issue is not indicative of an underlying problem with the marginal testlet model, but merely an issue with the current estimation program used. This issue should be resolved in further iterations of the program. Furthermore, the temporary fix of setting a_j to 1 did not influence the current results in a meaningful way.

If the marginal testlet model proves to be as adaptable and generally expandable as the conditional testlet model is today, we think that it is better suited as a model for estimating testlet dependencies in item response models. The benefits shown in this study, when compared to the conditional testlet model, demonstrate great potential. In order to investigate the usefulness and possible areas of application, further research is required. First, more simulation studies are necessary. The model should be tested when applied to different kinds of datasets. This includes using polytomous, instead of only dichotomous test items. Furthermore, different variations of items should be included, such as items without testlet effects. Importantly, the marginal testlet model should be used to obtain estimates on various real observed data-sets. Positive outcomes here would be a big step towards proving a general usefulness of the marginal testlet model. The underlying theoretical model could also be used in different fields than strictly psychometrics. It is our belief that the way in which the testlet model accounts for covariances within sets of variables has a wide range of possible applications, that are by no means limited to

psychometrics. We present a possible example in the context of agriculture. Imagine, we want to evaluate different kinds of crops. We could take the yields of the crops, planted on various plots of land over the course of multiple years as a measurement. In the context of the model, we could think of the different harvests as the test items and of the crops as the participants. The yields would thus make up the response observations. Over time, we would obtain measurements of the different crops planted on various plots. It is not unreasonable to assume that the different plots themselves also have an influence on the yield of the crops. This could be due to, for example, differences in soil composition, weather conditions etc. Therefore, every time a particular crop is planted on the same plot, the measurement of yield would in part rely on the constant factor of that plot of land. This is equivalent to the codependencies of testlet item responses of the testlet models discussed in this paper. We could, therefore, think of the plots as the testlet with the crop yield observations nested within them. Just as with the testlet models, the interdependencies of response observations within one testlet (plot of land) would make it possible to account for this influence, when estimating the general performance of the crops. Many more such examples could be imagined, where similar structures exist.

References

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

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

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

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: Theory and applications. *ETS Research Report Series,* 2002(1).

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

Poncet, P. (2012). modeest: Mode Estimation. R package version 2.1. https://CRAN.R-project.org/package=modeest

Appendix

Table 7

Mean Squared Error Between True Parameters and Estimated Posterior Means

		alpha		beta		theta	
con #	parameters:	conditional	marginal	conditional	marginal	conditional	marginal
1	1000, .05, 5	.0045	.0044	.0025	.0025	.0866	.0866
2	1000, .1, 10	.0045	.0045	.0026	.0026	.1081	.1081
3	1000, .01, 15	.0041	.0042	.0024	.0024	.0828	.0828
4	500, .1, 5	.0087	.0085	.0051	.0051	.0938	.0938
5	500, .01, 10	.0077	.0081	.0048	.0048	.0832	.0832
6	500, .05, 15	.0079	0(na)	.0051	.0047	.099	.099
7	200, .01, 5	.016	.0172	.0112	.0113	.0805	.0805
8	200, .05, 10	.0172	.0185	.0127	.0127	.0946	.0946
9	200, .1, 15	.0170	0(na)	.0121	.0113	.1203	.1203

Note. The MSE for the model parameters alpha (a), beta (b), and theta (θ). There are nearly no differences in MSE between the two models.

Table 8

Bias Between True Parameters and Posterior Estimate for the Model Parameters for the Conditional and Marginal Testlet Model

		alpha		beta		theta	
con #	parameters	conditional	marginal	conditional	marginal	conditional	marginal
2	1000, .1, 10	.0043	.0044	-.0026	-.0026	.0001	.0001
3	1000, .01, 15	.0039	.0041	.002	.002	.0001	.0001
4	500, .1, 5	.0084	.0082	.0001	.0001	.0001	.0001
5	500, .01, 10	.0075	.0079	.0029	.003	-.0001	-.0001
6	500, .05, 15	.0077	0(na)	.0033	.0029	-.0001	-.0001
7	200, .01, 5	.0158	.0169	.0008	.0005	.0001	.0001
8	200, .05, 10	.0168	.018	-.0013	-.0012	.0001	.0001
9	200, .1, 15	.0166	0(na)	.0028	.0018	.0001	.0001

Note. The bias for the model parameters alpha (a), beta (b), and theta (θ). The differences are very small between the estimations under the two models.