Behaviormetrika
Vol.37, No.2, 2010, 87–110

A SENSITIVITY ANALYSIS OF GIBBS SAMPLING FOR 3PNO IRT MODELS: EFFECTS OF PRIOR SPECIFICATIONS ON PARAMETER ESTIMATES

Yanyan Sheng*

The performance of the Gibbs sampling procedure for the three-parameter normal ogive (3PNO) IRT model was investigated using Monte Carlo simulations. Model parameters were estimated for tests with 10, 20, and 40 items and samples of 100, 300, 500, and 1000 examinees, where different actual values and prior specifications were considered for the item parameters. Summary statistics showed that this procedure was more affected by the choice of the prior distributions for the three-parameter model than the two-parameter model. For the 3PNO model, appropriate informative priors with relatively small spread should be adopted for the slope and intercept parameters to obtain more efficient and accurate MCMC estimates when sample sizes are not large and/or tests are not long enough. When it is not clear whether the prior information is appropriate, informative priors with small prior variances are not recommended.

1. Introduction

With current enhanced computational technology and the emergence of Markov chain Monte Carlo (MCMC) simulation techniques (e.g., Chib & Greenberg, 1995), the methodology for parameter estimation with item response theory (IRT) models has rapidly moved to a fully Bayesian approach. MCMC has been developed for various unidimensional item response models (e.g., Albert, 1992; Johnson & Albert, 1999; Patz & Junker, 1999a). However, it is not clear if the choice of the prior distributions for item parameters affects the performance of the procedure. This study is hence focusing on the impact of the prior specification on MCMC estimates of the model parameters.

For years, the standard methodology for parameter estimation in the IRT literature has been focusing on first calibrating item parameters by treating person parameters as missing data, and then using the estimated item parameters when making inference on persons. This marginal maximum likelihood (MML) estimation or the later emerged marginal Bayesian (MB; Mislevy, 1986) estimation, in spite of being less computational demanding compared with the MCMC estimation, suffers from a crucial drawback: it does not allow one to model the dependencies among parameters and sources of uncertainty (Patz & Junker, 1999a; Tsutakawa & Johnson, 1990; Tsutakawa & Soltys, 1988). Hence, MCMC methods present an alternative to MML or MB and offer promise for IRT parameter estimation.

MCMC is powerful for complicated models where the probabilities or expectations

Key Words and Phrases: item response theory, three-parameter normal ogive model, Gibbs sampling, sensitivity analysis, parameter estimation.

* Department of Educational Psychology & Special Education, Southern Illinois University Carbondale, Wham 223, Mailcode 4618, Carbondale, IL 62901, USA. E-mail: ysheng@siu.edu

are intractable by analytical methods or other numerical approaches. Its methods have been influential in modern Bayesian analyses where they are used to summarize the posterior distributions that arise in the context of the Bayesian prior-posterior framework (e.g., Carlin & Louis, 2000; Chib & Greenberg, 1995; Gelfand & Smith, 1990; Gelman, Carlin, Stern, & Rubin, 2003; Tanner & Wong, 1987). MCMC methods have proved useful in practically all aspects of Bayesian inference, such as parameter estimation and model comparisons. A key reason for the widespread interest in the MCMC method is that they are extremely general and flexible and hence can be used to sample univariate and multivariate distributions when other methods (e.g., MML) either fail or are difficult to implement. The rationale behind using MCMC has been given by Hastings (1970) and Metropolis and Ulam (1949), and the empirical advantage of using MCMC with IRT models has been provided by Wollack, Bolt, Cohen, and Lee (2002), among others.

One of the simplest MCMC algorithms is Gibbs sampling (Casella & George, 1992; Gelfand & Smith, 1990; Geman & Geman, 1984), which has very attractive theoretical properties of being geometrically ergodic and converging very quickly relative to other algorithms (e.g., Altman, Gill, & McDonald, 2004). The method is straightforward to implement when each full conditional distribution associated with a particular multivariate posterior distribution is a known distribution that is easy to sample. Gibbs sampling was first applied to IRT models by Albert (1992; see also Baker, 1998), who worked out the full conditional distributions for a two-parameter normal ogive (2PNO; e.g., Lord & Novick, 1968) model using the data augmentation idea of Tanner and Wong (1987), which helps accelerate convergence and hence improves the efficiency of the Gibbs sampler (e.g., Gelman et al., 2003). With no prior information concerning the item parameters, non-informative priors were adopted in their implementation so that inference was based solely on the data. Sahu (2002; see also Johnson & Albert, 1999) further generalized the approach to the three-parameter normal ogive (3PNO; Lord, 1980) model. Recent studies have utilized this Bayesian procedure for 3PNO models (e.g., Béguin & Glas, 2001; Glas & Meijer, 2003). In these applications, the investigators switched to using informative priors for the item parameters. It is understood that with an additional parameter, the 3PNO model is more complex than the 2PNO model, and hence suspected that the specification of prior distributions has a relatively larger effect on the Gibbs sampling procedure for the 3PNO model.

In the IRT literature, it is well accepted that large sample sizes are needed to estimate model parameters (e.g., Swaminathan & Gifford, 1983), which makes IRT less applicable in small sample situations. However, a major characteristic of Bayesian analysis is its ability to construct prior distributions for model parameters, which has been particularly useful for small datasets. Imposing an informative prior distribution on any parameter has the effect of constraining its estimate from assuming unreasonable values. Indeed, studies on the performance of the MB or similar Bayesian estimation have shown that by incorporating prior information about item parameters, model parameters can be estimated more accurately with smaller sample sizes (Lim & Drasgow, 1990; Mislevy, 1986; Swaminathan & Gifford, 1983, 1985, 1986).

In particular, for the three-parameter model, the use of informative priors for item parameters, especially those in the form of judge's ratings, has been found to substantially improve parameter estimation in small sample situations (Swaminathan & Gifford, 1986; Swaminathan et al., 2003).

With respect to informative priors, the variance of a prior distribution plays a critical role in estimating parameters, as small prior variances result in Bayesian estimates that are closer to the mean of the prior distribution than do larger variances. Hence, if the prior distribution is appropriately specified, the estimates will be less likely to take unreasonable values. Research investigating parameter recovery for two-parameter IRT models using the MB estimation suggests that prior variances have little effect on accurate estimates of model parameters in large datasets (Baker, 1990). However, they do make a difference in small datasets, where larger prior variances lead to a less accurate parameter estimation (Harwell & Janosky, 1991). It has to be emphasized that this property relies on the specification of appropriate prior distributions, as inappropriate choices of priors can lead to biased estimates and consequently incorrect inferences (Mislevy, 1986). However, little empirical evidence is available on whether prior variances play a critical role in reducing bias introduced by misspecificed priors on the quality of item parameter estimation.

In view of the above, this study is hence to investigate the impact of the choice of informative and non-informative prior distributions on the 3PNO model parameter estimates using Gibbs sampling, the relationship between sample sizes and the specification of prior distributions on the accuracy with which model parameters are estimated, as well as the role of prior variances in the quality of MCMC estimates. The results should help define limits on the appropriate use of the Gibbs sampler in accurate estimation of the IRT model parameters.

It has to be noted that when the full conditional distributions cannot be obtained in closed form, more complicated MCMC procedures have to be adopted. For example, Patz and Junker (1999a, 1999b) adopted Metropolis-Hastings within Gibbs (Chib & Greenberg, 1995) for the two-parameter logistic (2PL) and the three-parameter logistic (3PL) models. As Gibbs sampling is relatively easier to implement compared with other MCMC algorithms (Gelman et al., 2003), and the logistic and normal ogive models are identical in model fit or parameter estimates (Birnbaum, 1968; Embretson & Reise, 2000), the MCMC procedures for logistic models are not considered in this study.

The remainder of the paper is organized as follows. The 3PNO IRT model, as well as the 2PNO model, is briefly outlined in Section 2, with a description of the Gibbs sampler and prior specifications for the model parameters. Section 3 presents a simulation study on the sensitivity analysis with the 3PNO model where sample sizes, test lengths, choices of the prior distributions for item parameters, and distances between item parameters and their prior means are taken into consideration. For the purpose of comparisons, a similar simulation study is presented in Section 4 pertaining to the sensitivity analysis with the simpler 2PNO model. Finally, a few summary remarks are provided in Section 5.

2. Model and the Gibbs sampling procedure

The unidimensional IRT model provides a fundamental framework in modeling the person-item interaction by assuming one ability dimension. The conventional 2PNO model and the 3PNO model are defined as

$$P(y_{ij}=1|\theta_i,\alpha_j,\beta_j)=\Phi(\alpha_j\theta_j-\beta_j)=\int_{-\infty}^{\alpha_j\theta_j-\beta_j}\frac{1}{\sqrt{2\pi}}e^{\frac{-t^2}{2}}dt, \tag{1}$$

and

$$\begin{aligned}P(y_{ij}&=1|\theta_i,\alpha_j,\beta_j,\gamma_j)=\gamma_j+(1-\gamma_j)\Phi(\alpha_j\theta_i-\beta_j)\\&=\Phi(\alpha_j\theta_i-\beta_j)+\gamma_j(1-\Phi(\alpha_j\theta_i-\beta_j)),\quad 0\le\gamma_j<1\end{aligned} \tag{2}$$

respectively, where α_j is the slope parameter describing the item discrimination, β_j is the intercept parameter that is associated with item difficulty β_j^* such that $\beta_j=\alpha_j\beta_j^*$, γ_j, is the pseudo-chance-level parameter, and θ_i is a scalar ability parameter. The three-parameter model is theoretically more appealing and is applicable to a variety of testing situations where the one- or two-parameter models may be inadequate. However, it leads to estimation problems especially when using the maximum likelihood method (cf. Embretson & Reise, 2000).

Gibbs sampling is one member of a class of MCMC methods that may be used to obtain item and person parameter estimates simultaneously. It is applicable when the joint distribution is not known explicitly, but the conditional distribution of each variable is known. Its general underlying strategy is to iteratively sample each item parameter, e.g., ξ_j, where $\xi_j=(\alpha_j,\beta_j)'$, and person parameter, θ_i, from its own posterior distribution, conditional on the sampled values of all other person and item parameters, with starting values $\xi_j^{(0)}$ and $\theta_i^{(0)}$. This iterative process continues for a sufficient number of samples after the posterior distributions converge to stationary distributions (a phase referred to as burn-in). As with standard Monte Carlo, the posterior means of all the samples collected after burn-in are considered as estimates of the true parameters ξ_j and θ_i.

For a detailed illustration of the Gibbs sampling procedure for the 3PNO model, see Béguin and Glas (2001) or Glas and Meijer (2003).This procedure was implemented using MATLAB (Mathworks, Inc., 2005) in this study.

A prior distribution has to be imposed on each model parameter to implement the Gibbs sampler. In the literature, before the fully Bayesian method was developed for IRT models in the 1990s, the MB or similar Bayesian estimation has been the standard estimation method, in which the logistic form is mathematically easier with respect to numerical integration involved in the procedure. Hence, specifications of the prior distributions for item parameters in the logistic IRT model as well as their effect have been well studied (e.g., Gifford & Swaminathan, 1990; Harwell & Janosky, 1991; Lim & Drasgow, 1990; Mislevy, 1986; Swaminathan & Gifford, 1982, 1985, 1986; Swaminathan et al., 2003). The choice of prior distributions for a particular parameter is arbitrary. However, only a limited number have been used in practice and investigated in the simulation studies. With prior information, a lognormal prior

is usually chosen for α_j as the theory assumes an increasing probability function, and normal priors with specific location and scale parameters are considered for θ_i and β_j in the logistic model (Baker, 1990; Harwell & Janosky, 1991; Lim & Drasgow, 1990; Yen, 1987). Moreover, since γ_j ranges from 0 to 1, a beta prior is recommended (Swaminathan & Gifford, 1986).

As logistic and normal ogive models are essentially indistinguishable given proper scaling, their prior specifications are not supposed to differ much. Indeed, for normal ogive models, a standard normal prior is commonly adopted for the person parameter θ_i (Albert, 1992; Baker, 1998; Béguin & Glas, 2001; Glas & Meijer, 2003; Johnson & Albert, 1999) so that unique scaling is ensured and hence a particular identification problem in the 2PNO or 3PNO model can be resolved (see e.g. Albert 1992 for a description of the problem). With prior information, a beta prior distribution is usually assumed for γ_j and a normal prior is assumed for β_j (Béguin & Glas, 2001; Glas & Meijer, 2003; Johnson & Albert, 1999; Sahu, 2002). For α_j, Béguin and Glas (2001) and Glas and Meijer (2003) adopted a lognormal prior in their applications of MCMC for the 3PNO model. However, this prior specification would fail to lead to closed-form full conditional distributions, which makes it impossible to implement the Gibbs sampler. For this particular reason, Sahu (2002) considered a truncated normal prior for α_j, which is "advantageous to work on because of conjugacy" (p.220). Conjugate priors, in spite of being occasionally criticized as being too restrictive, are computationally convenient, interpretable as additional data, and flexible enough in the highly parameterized model setting (e.g., Berger, 1985; Carlin, 1996; Carlin & Louis, 2000; Gelman et al., 2003).

Therefore, due to the reasons that they result in closed-form full conditional distributions and that the focus of this study is on the Gibbs sampler, certain conjugate priors are considered in this study for item parameters in the 2PNO and 3PNO models when there is prior information so that $\alpha_j \sim N_{(0,\infty)}(\mu_\alpha, \sigma_\alpha^2)$ (a truncated normal), $\beta_j \sim N(\mu_\beta, \sigma_\beta^2)$, and $\gamma_j \sim Beta(a, b)$. Values of the hyperparameters in the prior distributions are specified differently given the prior information. It is noted that smaller values of σ_α^2 and σ_β^2, while larger values of a and b lead to more precise prior information. In practice, a and b can be chosen in a way that $E(\gamma_j) = \dfrac{a}{a+b}$ is some pre-specified value, such as $1/M$, where M denotes the number of alternatives in a multiple-choice test (Swaminathan & Gifford, 1986). Alternatively, it is possible to estimate the hyperparameters if they are unknown, in which case, suitable prior distributions have to be specified for them (see e.g., Swaminathan & Gifford, 1982, 1985, 1986). This approach, however, has not been pursued in this study.

3. Simulation 1

3.1 Method

To investigate the impact of the choice of the prior distributions on the 3PNO model parameter estimates using Gibbs sampling, a simulation study was carried out,

where four factors were manipulated, namely, sample sizes, test lengths, specifications of prior distributions for the item parameters, and distances between true item parameters and their prior means. Item responses for n individuals ($n = 100$, 300, 500, 1000) and k items ($k = 10$, 20, 40) were generated according to the 3PNO model, as defined in (2), where ability parameters were generated as samples from a standard normal distribution, and item parameters were generated from uniform distributions as described on the following page.

When implementing Sahu's (2002) Gibbs sampling procedure, three prior distributions were considered for γ_j (prior_γ). They were (1) $Beta(1,1)$, a flat prior; (2) $Beta(2,7)$, with a mean of approximately 0.22 (corresponding to an examinee's choosing one of the four options in a multiple choice item randomly) and a standard deviation of about 0.131; and (3) $Beta(5,17)$, following Zimowski, Muraki, Mislevy and Bock (2003), which has a similar mean as $Beta(2,7)$ but a smaller variability. In addition, four ways of setting the prior distributions for α_j (prior_α) or β_j (prior_β) were considered such that each had

1. a non-informative uniform prior, i.e., $p(\alpha_j) > 0$ or $p(\beta_j) \propto 1$;
2. a non-informative proper prior with a large prior variance, i.e., $\alpha_j \sim N_{(0,\infty)}(0, 10^{10})$ or $\beta_j \sim N(0, 10^{10})$;
3. an informative prior with σ_α^2 or σ_β^2 being 4, i.e., $\alpha_j \sim N_{(0,\infty)}(0,4)$ or $\beta_j \sim N(0,4)$;
4. an informative prior with a much smaller σ_α^2 or σ_β^2, i.e., $\alpha_j \sim N_{(0,\infty)}(0,1)$ or $\beta_j \sim N(0,1)$.

It's noted that only uniform and conjugate normal priors were considered for α_j and β_j. In addition, the last three prior specifications differ only in the value of the scale hyperparameters, as a major focus of this study was on the effect of prior variances on the quality of posterior estimates.The effect of prior specification was investigated independently for each item parameter, holding prior densities for the other parameters constant so that $\alpha_j \sim N_{(0,\infty)}(0,1)$, $\beta_j \sim N(0,1)$, and/or $\gamma_j \sim Beta(1,1)$.

For generating item responses, the true item parameters were set to be about $-1, 0, 2, 4$ and 6 standard deviations away from the center location of their last prior specification illustrated previously. In particular, the prior density $\alpha_j \sim N_{(0,\infty)}(0,1)$ centers at about .80 with a standard deviation of about .603, $\beta_j \sim N(0,1)$ centers at 0 with a standard deviation of 1, and $\gamma_j \sim Beta(5,17)$ centers at about .23 with a standard deviation of about .087. The actual α_j values considered in this study were thus generated from five uniform distributions $U(.1,.3)$, $U(.7,.9)$, $U(1.9,2.1)$, $U(3.1,3.3)$ and $U(4.3,4.5)$, corresponding to the five distances relative to the prior mean. Similarly, the actual β_j values were generated from $U(-1.1,-.9)$, $U(-.1,.1)$, $U(1.9,2.1)$, $U(3.9,4.1)$, $U(5.9,6.1)$, and the actual γ_j values were generated from $U(.13,.15)$, $U(.22,.24)$, $U(.39,.41)$, $U(.57,.59)$, $U(.74,.76)$. It is noted that when the true parameters are -1, 0, or 2 standard deviations away from their prior mean, the corresponding informative prior specification is appropriate as the true parameters are within its range, whereas when the true parameters are 4 or 6 standard deviations away, the corresponding prior is misspecified. Again, the effect of distances

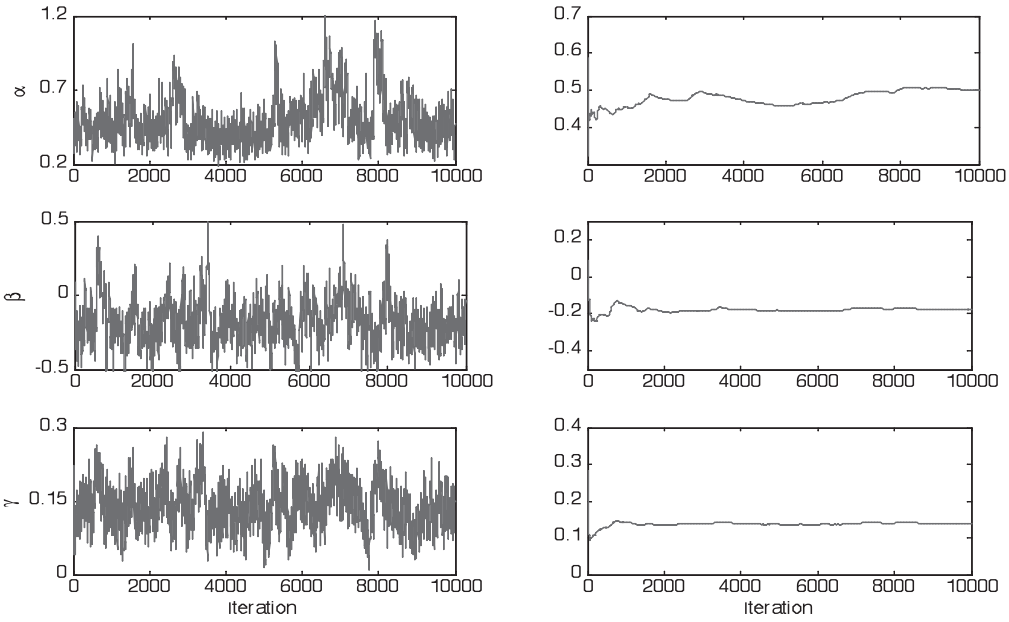

Figure 1: Trace plots (left) and running mean plots (right) of α, β, and γ for one item.

between true parameters and their prior means was investigated independently for each item parameter so that the others were generated as samples from $\alpha_j \sim U(0, 2)$, $\beta_j \sim U(-2, 2)$, and/or $\gamma_j \sim U(.05, .25)$.

The Gibbs sampling procedure was implemented where 10,000 to 50,000 iterations were obtained with the first half set as burn-in. Convergence was evaluated using the Gelman-Rubin R (Gelman & Rubin, 1992) statistic. The usual practice is using multiple Markov chains from different starting points. Alternatively, a single chain can be divided into sub-chains so that convergence is assessed by comparing the between and within sub-chain variance (Fox, 2007). Since a single chain is less wasteful in the number of iterations needed, the latter approach was adopted. For each Markov chain, the initial values were set to be $\alpha_j = 1$, $\beta_j = 0$, and $\gamma_j = .2$ for all items j and $\theta_i = 0$ for all persons i. After discarding the burn-in samples, the chain was then separated into five sub-chains of equal length and the R statistic was calculated following the procedure by Gelman and Rubin (1992). Convergence can also be monitored visually using time series graphs of the simulated sequence, such as the trace plot and the running mean plot shown in Figure 1 for one item. However, inspection of such plots has been criticized for being unreliable and unwieldy in the presence of a large number of model parameters (Gelman et al., 2003; Nylander, Wilgenbusch, Warren, & Swofford, 2008). The R statistic obtained from using a single chain was hence the major approach for assessing convergence in this study.

For each simulated scenario, 100 replications were conducted to avoid erroneous results in estimation due to sampling error. The accuracy of item parameter estimates was evaluated using the root mean square error ($RMSE$) and *bias*. Let τ denote the

true value of a parameter (α_j, β_j, or γ_j) and t_r its estimate in the rth replication ($r = 1, \ldots, R$). The *RMSE* is defined as

$$RMSE_\tau = \sqrt{\frac{\sum_{r=1}^{R}(t_r - \tau)^2}{R}}, \tag{3}$$

and the *bias* is defined as

$$BIAS_\tau = \frac{\sum_{r=1}^{R}(t_r - \tau)}{R}. \tag{4}$$

These quantities were averaged over items to provide summary indices.

3.2 Results

It is noted that when imposing an informative prior on α_j or β_j, all the chains converged to their stationary distributions within 10,000 to 50,000 iterations. Specifically, when the prior variance for α_j or β_j was small, e.g., 1, all the Markov chains reached stationarity with a run length of 10,000 to 30,000 iterations. It is when α_j or β_j was imposed with a larger prior variance, i.e., 4, that substantially more iterations were needed for the chains to converge. On the other hand, with non-informative priors (either uniform priors or conjugate normal priors with a large variance), stationarity was not reached even with a run length of 50,000 iterations when $k \leq 40$ and $n \leq 1000$, and hence their results are not reported. A close examination of the Gelman-Rubin R statistic at each replication indicates that with larger n and/or k, the Markov chains were more likely to reach stationarity. Hence, convergence can be improved by increasing the numbers of participants (n) and items (k).

The average Gelman-Rubin R statistics as well as the average *RMSE* and *bias* values for estimating α_j, β_j and γ_j are summarized in Tables 1, 2 and 3, respectively, where those associated with non-informative prior_α and prior_β are not displayed. A close examination of these values in Tables 1 and 2 leads to the following observations:

1). With a non-informative uniform prior specified for α_j or β_j, convergence was not observed in the Markov chains regardless of sample sizes, test lengths, or the actual values of these parameters. This is not surprising, as improper non-informative prior specifications lead to an improper joint posterior distribution for this model (Albert & Ghosh, 2000). On the other hand, imposing a proper non-informative prior_α or prior_β with a large variance did not seem to improve convergence.

2). Setting the prior variance for α_j or β_j to be relatively small helped with convergence, although it has to be noted that when β_j took values of $(3.9, 4.1)$ and $(5.9, 6.1)$, the Markov chains displayed relatively worse convergence within the simulated iterations using the second specification of prior_β where the prior variance was 4.

Table 1: Average Gelman-Rubin R, RMSE and bias for estimating the discrimination parameter (α) in the 3PNO model.

		R				*RMSE*				*bias*			
prior	true	$n=100$	$n=300$	$n=500$	$n=1000$	$n=100$	$n=300$	$n=500$	$n=1000$	$n=100$	$n=300$	$n=500$	$n=1000$
$k=10$													
$N_{(0,\infty)}(0,4)$	$(0.1,0.3)$	1.052	1.076	1.082	1.129	0.710	0.447	0.447	0.335	0.554	0.310	0.310	0.242
	$(0.7,0.9)$	1.050	1.072	1.078	1.088	0.831	0.666	0.666	0.518	0.580	0.423	0.423	0.264
	$(1.9,2.1)$	1.048	1.068	1.080	1.104	0.591	0.568	0.568	0.564	0.228	0.222	0.222	0.224
	$(3.1,3.3)$	1.045	1.093	1.081	1.077	0.749	0.584	0.581	0.465	−0.517	−0.232	−0.133	−0.036
	$(4.3,4.5)$	1.044	1.119	1.084	1.092	1.394	1.043	0.880	0.684	−1.299	−0.887	−0.692	−0.399
$N_{(0,\infty)}(0,1)$	$(0.1,0.3)$	1.033	1.098	1.071	1.124	0.395	0.279	0.279	0.247	0.326	0.212	0.212	0.186
	$(0.7,0.9)$	1.022	1.059	1.090	1.078	0.369	0.338	0.338	0.283	0.162	0.160	0.160	0.134
	$(1.9,2.1)$	1.017	1.042	1.075	1.068	0.562	0.419	0.419	0.323	−0.473	−0.319	−0.319	−0.143
	$(3.1,3.3)$	1.018	1.039	1.051	1.072	1.404	1.016	0.862	0.614	−1.379	−0.975	−0.805	−0.546
	$(4.3,4.5)$	1.016	1.037	1.057	1.077	2.394	1.920	1.664	1.283	−2.379	−1.899	−1.636	−1.239
$k=20$													
$N_{(0,\infty)}(0,4)$	$(0.1,0.3)$	1.041	1.124	1.111	1.129	0.622	0.460	0.435	0.295	0.470	0.315	0.288	0.200
	$(0.7,0.9)$	1.041	1.088	1.084	1.088	0.761	0.575	0.437	0.317	0.516	0.366	0.238	0.142
	$(1.9,2.1)$	1.042	1.089	1.100	1.104	0.613	0.565	0.517	0.410	0.321	0.265	0.192	0.169
	$(3.1,3.3)$	1.043	1.078	1.081	1.083	0.788	0.596	0.540	0.420	−0.570	−0.322	−0.255	−0.136
	$(4.3,4.5)$	1.043	1.096	1.081	1.077	1.481	1.006	0.842	0.650	−1.394	−0.841	−0.626	−0.417
$N_{(0,\infty)}(0,1)$	$(0.1,0.3)$	1.032	1.091	1.072	1.092	0.354	0.285	0.271	0.227	0.281	0.212	0.200	0.160
	$(0.7,0.9)$	1.021	1.049	1.058	1.078	0.334	0.305	0.255	0.213	0.132	0.151	0.115	0.077
	$(1.9,2.1)$	1.017	1.035	1.046	1.070	0.483	0.343	0.328	0.248	−0.395	−0.208	−0.154	−0.063
	$(3.1,3.3)$	1.017	1.037	1.049	1.076	1.461	1.097	0.945	0.702	−1.436	−1.062	−0.902	−0.651
	$(4.3,4.5)$	1.019	1.040	1.052	1.083	2.464	1.922	1.662	1.303	−2.452	−1.901	−1.635	−1.270
$k=40$													
$N_{(0,\infty)}(0,4)$	$(0.1,0.3)$	1.044	1.110	1.128	1.192	0.589	0.408	0.369	0.266	0.441	0.280	0.257	0.183
	$(0.7,0.9)$	1.038	1.060	1.069	1.069	0.738	0.542	0.406	0.235	0.524	0.323	0.209	0.112
	$(1.9,2.1)$	1.035	1.057	1.063	1.064	0.618	0.548	0.450	0.309	0.335	0.295	0.197	0.109
	$(3.1,3.3)$	1.041	1.072	1.085	1.081	0.888	0.684	0.561	0.434	−0.721	−0.470	−0.391	−0.251
	$(4.3,4.5)$	1.046	1.087	1.090	1.087	1.563	1.112	0.907	0.672	−1.470	−0.961	−0.729	−0.499
$N_{(0,\infty)}(0,1)$	$(0.1,0.3)$	1.033	1.105	1.087	1.174	0.313	0.262	0.255	0.210	0.241	0.188	0.184	0.148
	$(0.7,0.9)$	1.021	1.042	1.051	1.066	0.317	0.284	0.247	0.180	0.127	0.119	0.087	0.059
	$(1.9,2.1)$	1.016	1.028	1.040	1.044	0.491	0.326	0.299	0.228	−0.408	−0.174	−0.129	−0.065
	$(3.1,3.3)$	1.020	1.035	1.046	1.067	1.580	1.246	1.082	0.844	−1.559	−1.216	−1.053	−0.810
	$(4.3,4.5)$	1.021	1.040	1.053	1.076	2.559	2.059	1.785	1.422	−2.546	−2.040	−1.762	−1.395

Table 2: Average Gelman-Rubin R, RMSE and bias for estimating the intercept parameter (β) in the 3PNO model.

		R				*RMSE*				*bias*			
prior	true	$n=100$	$n=300$	$n=500$	$n=1000$	$n=100$	$n=300$	$n=500$	$n=1000$	$n=100$	$n=300$	$n=500$	$n=1000$
$k=10$													
$N(0,4)$	$(-1.1,-0.9)$	1.042	1.120	1.088	1.095	0.882	0.832	0.797	0.882	0.571	0.506	0.460	0.431
	$(-0.1,0.1)$	1.031	1.106	1.079	1.100	0.786	0.753	0.760	0.682	0.492	0.437	0.386	0.293
	$(1.9,2.1)$	1.043	1.095	1.100	1.123	0.563	0.568	0.531	0.633	0.043	0.090	0.072	0.144
	$(3.9,4.1)$	1.057	1.121	1.145	1.209	1.520	1.481	1.504	1.482	−1.482	−1.420	−1.433	−1.382
	$(5.9,6.1)$	1.055	1.113	1.178	1.322	3.512	3.448	3.374	3.295	−3.498	−3.424	−3.346	−3.243
$N(0,1)$	$(-1.1,-0.9)$	1.024	1.069	1.116	1.084	0.659	0.607	0.594	0.589	0.523	0.441	0.409	0.354
	$(-0.1,0.1)$	1.019	1.054	1.099	1.091	0.436	0.435	0.391	0.389	0.275	0.267	0.220	0.183
	$(1.9,2.1)$	1.019	1.058	1.090	1.104	0.647	0.551	0.494	0.457	−0.584	−0.485	−0.421	−0.349
	$(3.9,4.1)$	1.016	1.050	1.088	1.120	2.373	2.332	2.320	2.323	−2.359	−2.319	−2.306	−2.310
	$(5.9,6.1)$	1.018	1.051	1.087	1.123	4.376	4.304	4.264	4.262	−4.369	−4.297	−4.257	−4.255
$k=20$													
$N(0,4)$	$(-1.1,-0.9)$	1.040	1.088	1.077	1.086	0.850	0.858	0.862	0.627	0.547	0.475	0.472	0.328
	$(-0.1,0.1)$	1.038	1.070	1.095	1.070	0.779	0.640	0.610	0.476	0.477	0.357	0.291	0.200
	$(1.9,2.1)$	1.044	1.107	1.120	1.106	0.540	0.546	0.581	0.610	0.082	0.060	0.110	0.212
	$(3.9,4.1)$	1.057	1.133	1.181	1.403	1.562	1.484	1.449	1.426	−1.521	−1.433	−1.386	−1.320
	$(5.9,6.1)$	1.056	1.129	1.195	1.382	3.471	3.384	3.328	3.293	−3.457	−3.360	−3.302	−3.250
$N(0,1)$	$(-1.1,-0.9)$	1.024	1.059	1.096	1.089	0.635	0.594	0.592	0.522	0.480	0.403	0.386	0.299
	$(-0.1,0.1)$	1.022	1.045	1.063	1.071	0.438	0.376	0.368	0.295	0.259	0.222	0.185	0.134
	$(1.9,2.1)$	1.016	1.053	1.075	1.110	0.609	0.526	0.484	0.378	−0.539	−0.452	−0.383	−0.260
	$(3.9,4.1)$	1.018	1.053	1.090	1.120	2.411	2.332	2.288	2.295	−2.398	−2.321	−2.276	−2.282
	$(5.9,6.1)$	1.017	1.054	1.089	1.133	4.357	4.298	4.268	4.242	−4.351	−4.291	−4.261	−4.235
$k=40$													
$N(0,4)$	$(-1.1,-0.9)$	1.038	1.089	1.086	1.095	0.834	0.771	0.697	0.695	0.477	0.402	0.366	0.325
	$(-0.1,0.1)$	1.036	1.065	1.075	1.082	0.784	0.687	0.635	0.595	0.453	0.328	0.297	0.228
	$(1.9,2.1)$	1.043	1.110	1.119	1.119	0.534	0.521	0.543	0.508	0.109	0.082	0.154	0.140
	$(3.9,4.1)$	1.056	1.134	1.180	1.378	1.593	1.481	1.457	1.445	−1.551	−1.425	−1.388	−1.345
	$(5.9,6.1)$	1.057	1.135	1.181	1.437	3.507	3.344	3.310	3.206	−3.488	−3.320	−3.281	−3.156
$N(0,1)$	$(-1.1,-0.9)$	1.024	1.055	1.070	1.091	0.600	0.552	0.533	0.491	0.412	0.329	0.320	0.264
	$(-0.1,0.1)$	1.020	1.042	1.055	1.076	0.423	0.381	0.366	0.322	0.239	0.188	0.176	0.133
	$(1.9,2.1)$	1.017	1.046	1.069	1.106	0.591	0.500	0.413	0.348	−0.529	−0.427	−0.323	−0.231
	$(3.9,4.1)$	1.020	1.056	1.094	1.144	2.478	2.334	2.288	2.259	−2.465	−2.322	−2.275	−2.245
	$(5.9,6.1)$	1.020	1.053	1.091	1.138	4.430	4.290	4.260	4.227	−4.423	−4.284	−4.253	−4.219

Table 3: Average Gelman-Rubin R, RMSE and bias for estimating the pseudo-guessing parameter (γ) in the 3PNO model.

		R				$RMSE$				$bias$			
prior	true	$n=100$	$n=300$	$n=500$	$n=1000$	$n=100$	$n=300$	$n=500$	$n=1000$	$n=100$	$n=300$	$n=500$	$n=1000$
$k=10$													
$Beta(1,1)$	$(.13,.15)$	1.023	1.059	1.066	1.073	0.211	0.205	0.233	0.211	0.120	0.102	0.128	0.114
	$(.22,.24)$	1.026	1.067	1.067	1.077	0.227	0.189	0.187	0.192	0.118	0.075	0.058	0.073
	$(.39,.41)$	1.029	1.077	1.080	1.093	0.151	0.135	0.131	0.167	−0.002	−0.001	0.001	0.019
	$(.57,.59)$	1.039	1.075	1.087	1.173	0.151	0.146	0.143	0.138	−0.056	−0.052	−0.067	−0.051
	$(.74,.76)$	1.043	1.080	1.100	1.227	0.165	0.152	0.145	0.155	−0.122	−0.108	−0.108	−0.115
$Beta(2,7)$	$(.13,.15)$	1.015	1.036	1.064	1.064	0.077	0.073	0.086	0.081	0.051	0.040	0.054	0.046
	$(.22,.24)$	1.016	1.041	1.065	1.067	0.049	0.050	0.047	0.054	−0.008	−0.013	−0.017	−0.002
	$(.39,.41)$	1.018	1.052	1.084	1.061	0.152	0.136	0.131	0.124	−0.149	−0.133	−0.125	−0.114
	$(.57,.59)$	1.018	1.046	1.080	1.067	0.314	0.304	0.298	0.287	−0.313	−0.304	−0.297	−0.283
	$(.74,.76)$	1.014	1.030	1.045	1.059	0.482	0.479	0.477	0.471	−0.482	−0.478	−0.476	−0.470
$Beta(5,17)$	$(.13,.15)$	1.013	1.034	1.047	1.057	0.076	0.071	0.077	0.073	0.065	0.054	0.060	0.053
	$(.22,.24)$	1.012	1.032	1.049	1.058	0.027	0.031	0.029	0.035	−0.004	−0.007	−0.009	−0.001
	$(.39,.41)$	1.012	1.031	1.045	1.055	0.158	0.146	0.142	0.135	−0.157	−0.144	−0.140	−0.129
	$(.57,.59)$	1.012	1.024	1.037	1.062	0.334	0.330	0.327	0.323	−0.334	−0.330	−0.327	−0.322
	$(.74,.76)$	1.009	1.013	1.016	1.022	0.506	0.504	0.505	0.504	−0.506	−0.504	−0.505	−0.504
$k=20$													
$Beta(1,1)$	$(.13,.15)$	1.023	1.053	1.063	1.073	0.232	0.206	0.207	0.190	0.135	0.107	0.101	0.090
	$(.22,.24)$	1.024	1.066	1.072	1.079	0.210	0.193	0.198	0.165	0.097	0.085	0.083	0.057
	$(.39,.41)$	1.030	1.076	1.085	1.087	0.162	0.150	0.149	0.140	0.018	0.020	0.020	0.020
	$(.57,.59)$	1.037	1.085	1.081	1.194	0.155	0.144	0.138	0.132	−0.069	−0.058	−0.053	−0.055
	$(.74,.76)$	1.043	1.085	1.145	1.200	0.159	0.150	0.143	0.141	−0.118	−0.115	−0.107	−0.099
$Beta(2,7)$	$(.13,.15)$	1.014	1.033	1.047	1.056	0.080	0.077	0.076	0.072	0.051	0.044	0.042	0.037
	$(.22,.24)$	1.015	1.038	1.051	1.069	0.049	0.048	0.049	0.050	−0.012	−0.006	−0.006	−0.006
	$(.39,.41)$	1.016	1.045	1.068	1.072	0.145	0.130	0.125	0.113	−0.142	−0.123	−0.116	−0.093
	$(.57,.59)$	1.016	1.051	1.086	1.074	0.313	0.295	0.285	0.270	−0.312	−0.293	−0.281	−0.261
	$(.74,.76)$	1.012	1.030	1.046	1.067	0.482	0.475	0.470	0.462	−0.481	−0.475	−0.469	−0.459
$Beta(5,17)$	$(.13,.15)$	1.011	1.025	1.034	1.050	0.077	0.072	0.070	0.067	0.065	0.056	0.052	0.047
	$(.22,.24)$	1.011	1.026	1.041	1.051	0.029	0.030	0.032	0.031	−0.007	−0.003	−0.003	−0.003
	$(.39,.41)$	1.012	1.027	1.044	1.060	0.153	0.143	0.136	0.124	−0.152	−0.140	−0.132	−0.112
	$(.57,.59)$	1.010	1.020	1.032	1.076	0.333	0.326	0.322	0.312	−0.333	−0.326	−0.321	−0.309
	$(.74,.76)$	1.006	1.010	1.013	1.018	0.505	0.505	0.503	0.501	−0.505	−0.505	−0.503	−0.501
$k=40$													
$Beta(1,1)$	$(.13,.15)$	1.020	1.044	1.066	1.073	0.231	0.195	0.189	0.196	0.129	0.094	0.089	0.087
	$(.22,.24)$	1.023	1.056	1.072	1.073	0.212	0.186	0.172	0.171	0.111	0.075	0.068	0.065
	$(.39,.41)$	1.027	1.063	1.080	1.068	0.187	0.146	0.138	0.136	0.081	0.012	0.009	0.015
	$(.57,.59)$	1.035	1.088	1.092	1.182	0.147	0.135	0.133	0.126	−0.059	−0.055	−0.054	−0.041
	$(.74,.76)$	1.039	1.093	1.144	1.227	0.161	0.141	0.144	0.127	−0.122	−0.103	−0.105	−0.083
$Beta(2,7)$	$(.13,.15)$	1.013	1.029	1.038	1.056	0.079	0.074	0.072	0.071	0.048	0.040	0.038	0.034
	$(.22,.24)$	1.014	1.035	1.043	1.058	0.076	0.054	0.050	0.049	0.038	−0.007	−0.006	−0.002
	$(.39,.41)$	1.016	1.041	1.058	1.071	0.053	0.130	0.121	0.111	−0.012	−0.120	−0.106	−0.090
	$(.57,.59)$	1.013	1.051	1.083	1.085	0.313	0.289	0.276	0.254	−0.312	−0.286	−0.269	−0.235
	$(.74,.76)$	1.010	1.026	1.049	1.084	0.483	0.475	0.468	0.444	−0.483	−0.474	−0.466	−0.436
$Beta(5,17)$	$(.13,.15)$	1.010	1.022	1.028	1.042	0.074	0.069	0.067	0.064	0.061	0.052	0.048	0.043
	$(.22,.24)$	1.011	1.024	1.032	1.044	0.070	0.033	0.033	0.033	0.051	−0.004	−0.004	−0.001
	$(.39,.41)$	1.011	1.026	1.042	1.064	0.032	0.140	0.130	0.122	−0.007	−0.136	−0.122	−0.108
	$(.57,.59)$	1.008	1.020	1.035	1.092	0.333	0.323	0.317	0.300	−0.333	−0.323	−0.315	−0.293
	$(.74,.76)$	1.006	1.009	1.011	1.022	0.506	0.505	0.503	0.498	−0.506	−0.504	−0.503	−0.498

3). Increased sample sizes (n) consistently resulted in smaller average *RMSE* and *bias* for estimating α_j but not necessarily for β_j. On the other hand, increased test lengths (k) did not necessarily result in reduced *RMSE* or *bias* in estimating α_j or β_j.

4). Compared with the specification where σ_α^2 or σ_α^2 was 4, the last specification where σ_α^2 or σ_α^2 was 1 tended to result in smaller average *RMSE* and *bias* in estimating α_j or β_j when the true parameters were about -1, 0, or 2 standard deviations away from the prior mean of this specification, but it tended to result in consistently larger average *RMSE* and *bias* when the true parameters were about 4 or 6 standard deviations away. It is noted that the true parameters that are about 4 or 6 standard deviations away from the mean of $\alpha_j \sim N_{(0,\infty)}(0,1)$ or $\beta_j \sim N(0,1)$ are 2 or 3 standard deviations away from the mean of $\alpha_j \sim N_{(0,\infty)}(0,4)$ or $\beta_j \sim N(0,4)$.

5). When α_j or β_j took values of $(1.9, 2.1)$, which were about 1 standard deviation away from the prior mean of $\alpha_j \sim N_{(0,\infty)}(0,4)$ or $\beta_j \sim N(0,4)$, but 2 standard deviations away from the prior mean of $\alpha_j \sim N_{(0,\infty)}(0,1)$ or $\beta_j \sim N(0,1)$, the former prior specification resulted in positive average *bias*, whereas the latter resulted in negative average *bias*. Further, both specifications resulted in negative *bias* when the true parameters were greater than 3, indicating that they tended to underestimate α_j or β_j when these parameters were large (i.e., 2 or more standard deviations away from their prior means).

6). A comparison among the five ranges of true parameters indicates that $\beta_j \sim N(0,4)$ consistently resulted in smaller average *RMSE* and *bias* when β_j took values of $(1.9, 2.1)$, whereas $\beta_j \sim N(0,1)$ consistently resulted in smaller *RMSE* and *bias* when β_j were $(-.1, .1)$. On the other hand, $\alpha_j \sim N_{(0,\infty)}(0,4)$ tended to result in smaller *RMSE* when α_j were $(.1, .3)$ and smaller *bias* when the true parameters were $(1.9, 2.1)$, whereas $\alpha_j \sim N_{(0,\infty)}(0,1)$ tended to result in smaller *RMSE* and *bias* when α_j were $(.7, .9)$. Hence, these two prior specifications tended to perform better when α_j or β_j were within $\pm$ 1 standard deviation away from their prior means.

In general, sample sizes, test lengths, prior specifications, and distances between true parameters and prior means are the four key factors in the recovery of α_j or β_j. When sample sizes and/or test lengths are not large enough, less informative priors either fail to converge or require an extremely long chain to converge, which is computational demanding. Alternatively, when actual item parameters are not far from their prior means, informative prior_α and prior_β, especially those with relatively smaller spread, help accelerate convergence of Markov chains and consequently result in a more efficient and accurate estimation of the slope and intercept parameters. On the other hand, the non-convergence problem resulted from non-informative priors is less serious with increased sample sizes and/or test lengths, as the likelihood gets more informative. Hence, the choice of non-informative versus informative prior_α or prior_β does make a difference in the quality of α_j or β_j estimates. Since informative

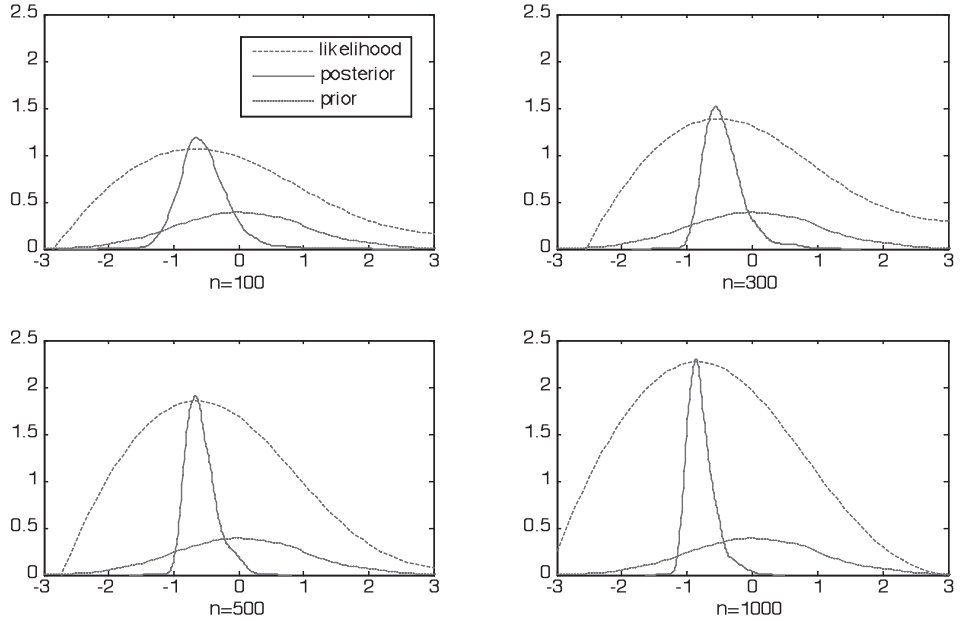

Figure 2: Prior, likelihood and posterior densities of β for an item with various sample sizes ($n = 100$, 300, 500, and 1000).

priors are recommended to be adopted for α_j or β_j to improve convergence, care has to be taken in noting whether the true parameters are within the range of the prior distribution. If not, the prior densities are inappropriate and tend to result in increased bias and error in estimation. It is noted that bias is relatively small when true parameters are close to (i.e., within $\pm$ 2 standard deviations away from) their prior means, but it increases when these parameters are farther away from their prior means.

For situations where stationarity has been reached for α_j or β_j, sample sizes play an important role in improving the accuracy of their posterior estimates. Indeed, Figure 2 displays the prior, likelihood and posterior density plots for an item parameter under various sample size conditions. It is clear from the figure that with small sample sizes, the posterior estimate is more affected by the prior specification, as less information can be drawn from the data (the likelihood). Additional data provide more information concerning the value of this parameter and hence the posterior estimate can be more accurate given an increased sample size.

Slightly different results were obtained regarding the recovery of the guessing or the chance level parameter (γ_j). Specifically, with either a non-informative or an informative prior$_\gamma$, the Gibbs sampler gave rise to Gelman-Rubin R statistics close to 1 regardless of sample sizes or test lengths, indicating that convergence had been reached within 30,000 iterations. This is different from the previous finding with estimating α_j or β_j. With respect to the posterior estimates of γ_j, increased n or k tended to, though not always, decrease the average *RMSE* and *bias*. A comparison

among the five ranges of true parameters indicates that with the flat prior, $Beta(1, 1)$, smaller average *RMSE* and *bias* were obtained when γ_j were about .4 or larger, where the posterior estimates tended to underestimate γ_j with a negative bias. On the other hand, with informative prior$_\gamma$, smaller *RMSE* and *bias* were obtained when γ_j were about .23, at which point, bias started to be negative. Among the two informative priors, $Beta(2, 7)$ and $Beta(5, 17)$, the latter has a smaller spread and performed better when γ_j were close to zero, but when γ_j were about .4 and larger, the former was preferred. This agrees with the finding with α_j or β_j in that $Beta(5, 17)$ was less appropriate than $Beta(2, 7)$ for the last three ranges of true parameters, as they were farther away from the mean of $Beta(5, 17)$ than that of $Beta(2, 7)$. Further, compared with the flat prior$_\gamma$, the two informative priors performed better in recovering γ_j when the true parameters were close to zero, but worse when the true parameters were away from zero. Comparing Table 3 with Table 1 or Table 2, one may note that various levels of prior specifications, actual values, sample sizes, or test lengths did not seem to result in *RMSE* or *bias* values for γ_j as large or different as those for α_j or β_j. However, this does not imply that these factors have considerably less effect on the accurate estimation of γ_j than the other item parameters. The smaller *RMSE* or *bias* values and hence differences are largely due to the fact that the full conditional distribution of γ_j is a beta distribution, which prevents their posterior samples from assuming values beyond the range of 0 and 1.

Hence, in general, relatively informative priors are suggested for item slope and intercept parameters in the 3PNO model particularly when sample sizes and/or test lengths are not very large. For the guessing parameter, both non-informative and informative priors can be adopted, with the latter being recommended for true parameters being close to zero, and the former being recommended for true parameters being relatively large. When specifying informative priors, one has to make sure that a small spread is only adopted when the true parameters are close to their prior means. If they are away from their prior means, small prior variances are more likely to result in biased MCMC estimates, as the posterior samples are restricted to take on values closer to the prior means. In this case, a relatively larger spread will need to be adopted to reduce bias and improve the accuracy of the posterior estimates. Furthermore, the bias introduced by prior means that are not close to actual item parameters may be lessened by increasing sample sizes so that the likelihood becomes more informative. One shall note that these findings on the use of informative priors agree with what is offered by the existing literature in Bayesian statistics (e.g., Gelman, Carlin, Stern, & Rubin, 2003).

4. Simulation 2

4.1 Method

For the purpose of comparisons, a similar simulation was carried out with the Gibbs sampling procedure for the 2PNO model, where four sample sizes ($n = 100$, 300, 500,

1000), three test lengths ($k = 10$, 20, 40), four prior distributions for α_j (prior$_\alpha$) or β_j (prior$_\beta$), as listed in the previous section, and five distances (-1, 0, 2, 4, and 6 standard deviations) between true item parameters and the center location of their last prior specifications were considered. Item responses were generated using the 2PNO model, as is shown in (1), where ability parameters were generated as samples from a standard normal distribution. For each item parameter under investigation, the other parameter was generated as samples from a uniform distribution so that $\alpha_j \sim U(0, 2)$ or $\beta_j \sim U(-2, 2)$, and was specified to assume a prior density of $\alpha_j \sim N_{(0,\infty)}(0, 1)$ or $\beta_j \sim N(0, 1)$. The initial values used in the Gibbs sampling procedure were $\alpha_j = 1$ and $\beta_j = 0$ for all items and $\theta_i = 0$ for all persons. 10,000 to 50,000 iterations of the Gibbs sampling were obtained with the first half set as burn-in and the posterior expectations of the Gibbs samples were obtained as the posterior estimates. With 100 replications, the accuracy of parameter estimates was evaluated using the *RMSE* and *bias*, as defined in (3) and (4). Their values were averaged over items to provide summary indices.

4.2 Results

For each implementation of the Gibbs sampler involved in this simulation study, convergence was assessed using the Gelman-Rubin R statistic via the single chain approach.The average Gelman-Rubin R statistic, together with the average *RMSE* and *bias* values for α_j and β_j are displayed in Tables 4 and 5. From the results, the following remarks can be made concerning the recovery of α_j and β_j in the 2PNO model using Gibbs sampling:

1). When β_j took values of $(3.9, 4.1)$ and $(5.9, 6.1)$, improper or proper non-informative prior$_\beta$ did not give rise to converged Markov chains within 50,000 iterations and hence their results are not reported in Table 4. Further, the average Gelman-Rubin R statistics for posterior samples of α_j with $n = 100$ and $k = 40$ were somewhat large when α_j took values of $(3.1, 3.3)$ and $(4.3, 4.5)$ (see Table 3), indicating that the Markov chains did not converge within 50,000 iterations. Stationarity had been reached under the other scenarios, although it has to be noted that with smaller sample sizes or test lengths, the Markov chains required more iterations to reach convergence.

2). Increased sample sizes (n), though not necessarily increased test lengths (k) tended to reduce bias and hence improve the accuracy of the estimation of α_j or β_j. Sample sizes seemed to play a bigger role. This is consistent with the findings from the 3PNO model.

3). Regardless of prior specifications, relatively smaller average *RMSE* and *bias* were obtained when α_j were about $(.1, .3)$ and β_j were about $(-.1, .1)$, i.e., when they were close to zero. This is slightly different from what was observed with the 3PNO model.

4). The two non-informative priors performed fairly similarly in estimating α_j or β_j.

It is noted that among them, the proper prior $\alpha_j \sim N_{(0,\infty)}(0, 10^{10})$ was slightly better in estimating α_j when the actual parameters were close to zero, and the improper prior $p(\alpha_j) > 0$ was slightly better when α_j were large. This general pattern was not observed for β_j given that convergence was not reached when β_j were large.

5). Among the two informative priors considered for α_j or β_j, the one with a smaller spread tended to perform better only when the true parameters were within the range of its distribution and sample sizes were not large. When sample sizes increased, non-informative priors tended to result in *RMSE* and *bias* closer to those resulted from appropriately specified informative priors. It is noted that the two informative priors tended to underestimate α_j with a negative bias when they were large. This was, however, not observed with the non-informative priors, indicating that they allow occasionally large values for α_j.

6). Comparing among the four prior specifications, we see that when sample sizes were small, informative priors were preferred provided that the true parameters were within the range of its distribution; when sample sizes increased, the difference between appropriate informative priors and non-informative priors became negligible. These results agree with the findings by Baker (1992) and Harwell and Janosky (1991) on the use of the MB estimation for the two-parameter model. Further, when true parameters were outside of the range of the prior density, the estimates tended to have a large bias and estimation error.

7). As was expected, α_j and β_j are estimated more accurately in the 2PNO model than in the more complex 3PNO model.

A comparison of Tables 1 and 2 vs. Tables 4 and 5 suggests that when true parameters of α_j or β_j were four or six standard deviations away from their prior means, the quality of parameter estimates in the 3PNO model was affected more than that in the 2PNO model. Furthermore, large values of β_j, e.g., $(3.9, 4.1)$ and $(5.9, 6.1)$, had a considerably greater effect on their posterior estimates in the 3PNO model than in the 2PNO model, particularly when sample sizes were large.

Consequently, similar yet slightly different conclusions can be drawn regarding the effects of sample sizes, test lengths, prior specifications, and distances between true parameters and prior means on parameter estimation in the 2PNO model. Specifically, non-informative priors can be specified for item slope and intercept parameters provided that the true intercept parameters are not very far from zero. With no prior information, specifying proper or improper prior densities does not result in much different estimates of the 2PNO model. When sample sizes are small and the true item parameters are close to the prior mean, a relatively small spread tends to help with convergence of the Markov chain and accuracy in estimating item parameters. However, when the true parameters are far away from the prior mean, small prior variances tend to result in biased and inaccurate MCMC estimates. In addition, with the same test length, the effect of misspecified priors can be offset by increasing sample sizes.

Table 4: Average Gelman−Rubin R, RMSE and bias for recovering the discrimination parameter (α) in the 2PNO model.

			R				*RMSE*				*bias*			
	prior	true	$n=100$	$n=300$	$n=500$	$n=1000$	$n=100$	$n=300$	$n=500$	$n=1000$	$n=100$	$n=300$	$n=500$	$n=1000$
$k=10$														
	Unif	$(0.1, 0.3)$	1.012	1.004	1.002	1.003	0.209	0.117	0.068	0.052	0.118	0.050	0.008	0.009
		$(0.7, 0.9)$	1.031	1.015	1.007	1.011	0.372	0.176	0.125	0.087	0.172	0.048	0.030	0.019
		$(1.9, 2.1)$	1.070	1.087	1.072	1.078	0.953	0.555	0.433	0.280	0.416	0.205	0.167	0.056
		$(3.1, 3.3)$	1.102	1.095	1.097	1.087	1.209	0.809	0.509	0.347	0.716	0.432	0.290	0.121
		$(4.3, 4.5)$	1.130	1.136	1.139	1.150	1.111	0.999	0.765	0.574	0.929	0.468	0.356	0.188
	$N_{(0,\infty)}(0, 10^{10})$	$(0.1, 0.3)$	1.012	1.003	1.002	1.003	0.207	0.117	0.068	0.053	0.118	0.050	0.009	0.009
		$(0.7, 0.9)$	1.034	1.014	1.008	1.010	0.367	0.178	0.126	0.087	0.166	0.048	0.031	0.019
		$(1.9, 2.1)$	1.080	1.086	1.073	1.079	0.935	0.534	0.408	0.285	0.469	0.184	0.158	0.063
		$(3.1, 3.3)$	1.111	1.098	1.093	1.085	1.413	0.821	0.514	0.346	0.834	0.417	0.292	0.118
		$(4.3, 4.5)$	1.110	1.128	1.132	1.150	1.341	0.892	0.735	0.568	1.274	0.296	0.302	0.155
	$N_{(0,\infty)}(0, 4)$	$(0.1, 0.3)$	1.003	1.003	1.002	1.004	0.193	0.115	0.068	0.052	0.107	0.048	0.008	0.008
		$(0.7, 0.9)$	1.010	1.009	1.008	1.009	0.329	0.170	0.124	0.086	0.130	0.041	0.026	0.017
		$(1.9, 2.1)$	1.030	1.054	1.057	1.087	0.557	0.479	0.359	0.275	0.168	0.120	0.095	0.034
		$(3.1, 3.3)$	1.037	1.055	1.056	1.071	0.662	0.473	0.331	0.285	−0.425	−0.160	−0.090	−0.070
		$(4.3, 4.5)$	1.046	1.082	1.101	1.136	1.134	0.719	0.570	0.454	−1.000	−0.473	−0.296	−0.156
	$N_{(0,\infty)}(0, 1)$	$(0.1, 0.3)$	1.002	1.003	1.002	1.003	0.166	0.110	0.067	0.052	0.087	0.043	0.005	0.007
		$(0.7, 0.9)$	1.007	1.008	1.007	1.009	0.257	0.155	0.117	0.084	0.046	0.017	0.014	0.011
		$(1.9, 2.1)$	1.013	1.027	1.033	1.056	0.423	0.307	0.256	0.220	−0.318	−0.136	−0.064	−0.052
		$(3.1, 3.3)$	1.017	1.031	1.041	1.046	1.221	0.795	0.603	0.449	−1.194	−0.749	−0.561	−0.396
		$(4.3, 4.5)$	1.020	1.039	1.055	1.080	2.094	1.470	1.176	0.844	−2.080	−1.446	−1.148	−0.801
$k=20$														
	Unif	$(0.1, 0.3)$	1.006	1.003	1.002	1.003	0.207	0.125	0.075	0.069	0.118	0.049	0.012	0.013
		$(0.7, 0.9)$	1.019	1.007	1.007	1.011	0.406	0.173	0.119	0.081	0.214	0.063	0.039	0.015
		$(1.9, 2.1)$	1.048	1.035	1.028	1.043	1.630	0.449	0.319	0.209	0.900	0.215	0.130	0.058
		$(3.1, 3.3)$	1.133	1.094	1.092	1.126	1.818	0.851	0.556	0.285	1.371	0.551	0.622	0.107
		$(4.3, 4.5)$	1.101	1.140	1.136	1.173	1.337	1.229	0.642	0.467	0.755	0.958	0.274	0.156
	$N_{(0,\infty)}(0, 10^{10})$	$(0.1, 0.3)$	1.006	1.003	1.002	1.003	0.205	0.127	0.075	0.069	0.116	0.050	0.013	0.013
		$(0.7, 0.9)$	1.020	1.007	1.008	1.009	0.403	0.173	0.119	0.081	0.210	0.062	0.039	0.016
		$(1.9, 2.1)$	1.066	1.037	1.028	1.039	1.309	0.464	0.319	0.207	0.906	0.218	0.130	0.055
		$(3.1, 3.3)$	1.112	1.103	1.092	1.139	1.755	0.844	0.543	0.296	1.246	0.539	0.321	0.121
		$(4.3, 4.5)$	1.092	1.121	1.134	1.176	1.539	1.060	0.716	0.394	1.381	0.471	0.369	0.114
	$N_{(0,\infty)}(0, 4)$	$(0.1, 0.3)$	1.003	1.003	1.002	1.003	0.179	0.122	0.074	0.068	0.092	0.045	0.011	0.012
		$(0.7, 0.9)$	1.008	1.007	1.007	1.010	0.313	0.163	0.114	0.080	0.116	0.045	0.029	0.011
		$(1.9, 2.1)$	1.024	1.030	1.025	1.039	0.639	0.344	0.273	0.191	0.329	0.103	0.074	0.031
		$(3.1, 3.3)$	1.034	1.055	1.065	1.105	0.766	0.499	0.383	0.277	−0.563	−0.307	−0.216	−0.153
		$(4.3, 4.5)$	1.045	1.079	1.096	1.173	1.252	0.792	0.629	0.450	−1.129	−0.608	−0.469	−0.263
	$N_{(0,\infty)}(0, 1)$	$(0.1, 0.3)$	1.002	1.003	1.002	1.003	0.148	0.112	0.072	0.067	0.062	0.035	0.005	0.010
		$(0.7, 0.9)$	1.006	1.006	1.006	1.008	0.236	0.144	0.107	0.078	0.004	0.006	0.005	0.000
		$(1.9, 2.1)$	1.013	1.018	1.019	1.035	0.367	0.263	0.222	0.177	−0.208	−0.105	−0.057	−0.035
		$(3.1, 3.3)$	1.018	1.029	1.038	1.074	1.350	0.954	0.770	0.569	−1.324	−0.921	−0.737	−0.539
		$(4.3, 4.5)$	1.020	1.040	1.053	1.111	2.225	1.624	1.361	0.974	−2.212	−1.601	−1.338	−0.940
$k=40$														
	Unif	$(0.1, 0.3)$	1.004	1.003	1.003	1.003	0.243	0.110	0.083	0.055	0.161	0.046	0.023	0.007
		$(0.7, 0.9)$	1.015	1.009	1.012	1.013	0.461	0.170	0.121	0.084	0.329	0.078	0.048	0.030
		$(1.9, 2.1)$	1.042	1.028	1.038	1.038	1.425	0.465	0.304	0.183	1.130	0.293	0.174	0.079
		$(3.1, 3.3)$	1.421	1.108	1.120	1.148	n.c.	1.084	0.657	0.351	n.c.	0.866	0.460	0.208
		$(4.3, 4.5)$	1.705	1.105	1.144	1.204	n.c.	1.218	0.950	0.412	n.c.	0.912	0.676	0.123
	$N_{(0,\infty)}(0, 10^{10})$	$(0.1, 0.3)$	1.005	1.003	1.004	1.003	0.243	0.110	0.083	0.055	0.161	0.046	0.023	0.007
		$(0.7, 0.9)$	1.017	1.009	1.014	1.012	0.463	0.170	0.121	0.084	0.331	0.079	0.048	0.030
		$(1.9, 2.1)$	1.042	1.029	1.042	1.039	1.485	0.463	0.304	0.182	1.165	0.293	0.176	0.079
		$(3.1, 3.3)$	1.420	1.111	1.108	1.154	n.c.	1.055	0.664	0.357	n.c.	0.832	0.459	0.198
		$(4.3, 4.5)$	1.740	1.147	1.141	1.186	n.c.	2.082	0.975	0.460	n.c.	1.276	0.681	0.400
	$N_{(0,\infty)}(0, 4)$	$(0.1, 0.3)$	1.003	1.002	1.003	1.003	0.180	0.103	0.081	0.054	0.100	0.037	0.019	0.005
		$(0.7, 0.9)$	1.007	1.007	1.011	1.013	0.276	0.150	0.113	0.080	0.121	0.044	0.031	0.021
		$(1.9, 2.1)$	1.018	1.024	1.034	1.039	0.584	0.335	0.250	0.166	0.291	0.145	0.103	0.047
		$(3.1, 3.3)$	1.032	1.048	1.059	1.115	0.866	0.548	0.462	0.317	−0.702	−0.438	−0.348	−0.223
		$(4.3, 4.5)$	1.045	1.077	1.091	1.184	1.357	0.881	0.710	0.490	−1.245	−0.741	−0.573	−0.352
	$N_{(0,\infty)}(0, 1)$	$(0.1, 0.3)$	1.002	1.002	1.003	1.003	0.134	0.091	0.076	0.053	0.052	0.019	0.008	0.000
		$(0.7, 0.9)$	1.005	1.006	1.011	1.012	0.200	0.131	0.102	0.075	−0.042	−0.022	−0.011	−0.001
		$(1.9, 2.1)$	1.011	1.018	1.031	1.035	0.409	0.254	0.200	0.149	−0.288	−0.114	−0.063	−0.038
		$(3.1, 3.3)$	1.018	1.031	1.039	1.087	1.495	1.127	0.965	0.730	−1.472	−1.107	−0.942	−0.710
		$(4.3, 4.5)$	1.023	1.042	1.056	1.126	2.362	1.800	1.541	1.164	−2.349	−1.781	−1.521	−1.141

Table 5: Average Gelman-Rubin R, RMSE and bias for recovering the intercept parameter (β) in the 2PNO model.

		R				*RMSE*				*bias*			
prior	true	$n=100$	$n=300$	$n=500$	$n=1000$	$n=100$	$n=300$	$n=500$	$n=1000$	$n=100$	$n=300$	$n=500$	$n=1000$
$k=10$													
Unif	$(-1.1,-0.9)$	1.010	1.012	1.011	1.018	0.268	0.154	0.110	0.083	−0.057	−0.025	0.000	−0.015
	$(-0.1,0.1)$	1.006	1.007	1.007	1.009	0.191	0.106	0.080	0.066	−0.007	−0.005	0.006	−0.015
	$(1.9,2.1)$	1.058	1.043	1.041	1.060	2.173	0.374	0.243	0.159	0.531	0.136	0.070	0.002
$N(0,10^{10})$	$(-1.1,-0.9)$	1.009	1.011	1.014	1.020	0.271	0.157	0.108	0.084	−0.058	−0.026	0.000	−0.015
	$(-0.1,0.1)$	1.006	1.007	1.007	1.010	0.192	0.105	0.080	0.065	−0.007	−0.006	0.007	−0.015
	$(1.9,2.1)$	1.051	1.040	1.035	1.052	2.565	0.367	0.237	0.160	0.590	0.134	0.069	0.000
$N(0,4)$	$(-1.1,-0.9)$	1.009	1.011	1.011	1.018	0.252	0.151	0.108	0.083	−0.045	−0.023	0.001	−0.015
	$(-0.1,0.1)$	1.006	1.007	1.007	1.010	0.188	0.106	0.081	0.066	−0.015	−0.009	0.003	−0.016
	$(1.9,2.1)$	1.029	1.034	1.033	1.048	0.429	0.299	0.216	0.154	0.108	0.075	0.039	−0.010
	$(3.9,4.1)$	1.046	1.136	1.164	1.205	0.859	0.674	0.671	0.559	−0.718	−0.441	−0.391	−0.191
	$(5.9,6.1)$	1.050	1.113	1.150	1.219	2.410	2.110	1.931	1.841	−2.406	−2.094	−1.915	−1.807
$N(0,1)$	$(-1.1,-0.9)$	1.007	1.009	1.009	1.016	0.222	0.144	0.103	0.081	−0.008	−0.016	0.005	−0.012
	$(-0.1,0.1)$	1.005	1.006	1.006	1.010	0.181	0.106	0.081	0.067	−0.021	−0.015	−0.002	−0.018
	$(1.9,2.1)$	1.010	1.018	1.022	1.041	0.314	0.224	0.187	0.148	−0.147	−0.043	−0.031	−0.046
	$(3.9,4.1)$	1.014	1.034	1.060	1.112	1.563	1.237	1.137	0.908	−1.546	−1.208	−1.100	−0.866
	$(5.9,6.1)$	1.012	1.029	1.049	1.088	3.399	3.034	2.862	2.683	−3.399	−3.032	−2.861	−2.681
$k=20$													
Unif	$(-1.1,-0.9)$	1.008	1.009	1.012	1.012	0.211	0.142	0.089	0.073	−0.004	−0.040	−0.008	−0.003
	$(-0.1,0.1)$	1.007	1.006	1.009	1.010	0.194	0.116	0.078	0.058	−0.002	−0.014	−0.001	−0.003
	$(1.9,2.1)$	1.057	1.027	1.031	1.032	2.807	0.505	0.210	0.141	0.634	0.112	0.041	0.007
$N(0,10^{10})$	$(-1.1,-0.9)$	1.008	1.009	1.012	1.011	0.209	0.141	0.090	0.073	−0.002	−0.040	−0.007	−0.003
	$(-0.1,0.1)$	1.006	1.007	1.010	1.010	0.194	0.116	0.079	0.058	−0.001	−0.014	0.000	−0.003
	$(1.9,2.1)$	1.070	1.056	1.032	1.030	2.454	0.622	0.206	0.140	0.589	0.128	0.039	0.006
$N(0,4)$	$(-1.1,-0.9)$	1.007	1.009	1.013	1.012	0.201	0.138	0.089	0.073	−0.006	−0.040	−0.009	−0.003
	$(-0.1,0.1)$	1.005	1.006	1.009	1.010	0.188	0.114	0.078	0.059	−0.017	−0.020	−0.004	−0.004
	$(1.9,2.1)$	1.018	1.022	1.028	1.028	0.441	0.270	0.196	0.138	0.089	0.038	0.019	−0.002
	$(3.9,4.1)$	1.046	1.116	1.150	1.179	0.832	0.623	0.545	0.524	−0.712	−0.418	−0.310	−0.088
	$(5.9,6.1)$	1.047	1.116	1.151	1.215	2.439	2.085	1.956	1.795	−2.434	−2.074	−1.941	−1.761
$N(0,1)$	$(-1.1,-0.9)$	1.006	1.008	1.012	1.013	0.181	0.131	0.086	0.073	0.014	−0.038	−0.010	−0.004
	$(-0.1,0.1)$	1.006	1.005	1.010	1.010	0.179	0.116	0.078	0.059	−0.035	−0.031	−0.012	−0.009
	$(1.9,2.1)$	1.010	1.015	1.024	1.027	0.315	0.219	0.179	0.135	−0.141	−0.059	−0.039	−0.033
	$(3.9,4.1)$	1.015	1.036	1.056	1.098	1.583	1.197	1.043	0.826	−1.568	−1.178	−1.022	−0.800
	$(5.9,6.1)$	1.014	1.030	1.045	1.086	3.440	3.029	2.871	2.667	−3.439	−3.028	−2.870	−2.664
$k=40$													
Unif	$(-1.1,-0.9)$	1.010	1.009	1.013	1.013	0.243	0.139	0.102	0.072	−0.047	−0.016	−0.010	−0.008
	$(-0.1,0.1)$	1.008	1.009	1.011	1.013	0.203	0.116	0.082	0.058	−0.023	−0.012	−0.003	−0.006
	$(1.9,2.1)$	1.046	1.018	1.027	1.026	1.703	0.260	0.190	0.128	0.408	0.059	0.046	0.016
$N(0,10^{10})$	$(-1.1,-0.9)$	1.009	1.009	1.014	1.013	0.245	0.138	0.103	0.072	−0.049	−0.015	−0.012	−0.007
	$(-0.1,0.1)$	1.008	1.008	1.012	1.012	0.204	0.116	0.083	0.057	−0.025	−0.011	−0.004	−0.005
	$(1.9,2.1)$	1.060	1.019	1.028	1.025	2.489	0.262	0.188	0.127	0.526	0.061	0.044	0.018
$N(0,4)$	$(-1.1,-0.9)$	1.008	1.008	1.014	1.013	0.236	0.137	0.102	0.072	−0.063	−0.024	−0.016	−0.009
	$(-0.1,0.1)$	1.008	1.008	1.012	1.012	0.200	0.114	0.082	0.058	−0.052	−0.023	−0.010	−0.008
	$(1.9,2.1)$	1.017	1.017	1.027	1.026	0.423	0.235	0.176	0.124	0.043	0.015	0.020	0.005
	$(3.9,4.1)$	1.048	1.110	1.137	1.207	0.866	0.601	0.536	0.513	−0.751	−0.422	−0.274	−0.180
	$(5.9,6.1)$	1.048	1.123	1.149	1.242	2.491	2.111	1.968	1.794	−2.487	−2.101	−1.953	−1.765
$N(0,1)$	$(-1.1,-0.9)$	1.006	1.007	1.011	1.014	0.209	0.133	0.100	0.073	−0.051	−0.033	−0.024	−0.016
	$(-0.1,0.1)$	1.005	1.006	1.010	1.013	0.198	0.119	0.083	0.060	−0.080	−0.044	−0.025	−0.018
	$(1.9,2.1)$	1.009	1.013	1.021	1.024	0.334	0.216	0.163	0.122	−0.181	−0.078	−0.040	−0.029
	$(3.9,4.1)$	1.020	1.036	1.053	1.093	1.958	1.206	1.028	0.834	−1.933	−1.190	−1.009	−0.810
	$(5.9,6.1)$	1.019	1.032	1.049	1.085	3.743	3.054	2.877	2.663	−3.742	−3.053	−2.876	−2.661

5. Discussion

In summary, the simulation studies on the performance of Gibbs sampling procedures do suggest that with the introduction of the pseudo-chance level parameter, the 3PNO model is more affected by the choice of the prior distributions, especially those for the slope (or discrimination) and intercept parameters, than the 2PNO model using the Gibbs sampler. For the 2PNO model, parameter estimation is not sensitive

to the choice of the prior distributions when sample sizes are relatively large (e.g., 1000 or more subjects). To make prior densities to be not informative, one can choose uniform priors or set prior variances to be extremely large. These priors, however, create problem in convergence when the intercept parameters are outside of $(-4, 4)$ regardless of sample sizes, and are therefore not recommended in such situations for the 2PNO model. On the other hand, nonconvergence poses a critical issue for the 3PNO model if non-informative prior densities are adopted for the item slope and intercept parameters even for large datasets. Improper priors have to be avoided for these parameters, as the 3PNO model, defined in (2), can be viewed as a mixture model with two components, i.e., 1 and $\Phi(\alpha_j, \theta_i - \beta_j)$, so the probability of an observation coming from a component is γ_j or $1 - \gamma_j$. It is said that the prior for the non-component parameter of the mixture (γ_j in this context) can be chosen in a typical fashion (Richardson & Green, 1997), and further that improper priors for component specific parameters (e.g., α_j, β_j) are not recommended, as the joint posterior distribution is not defined and parameter estimates are likely to be unstable (see Diebolt & Robert, 1994; Mengersen & Robert, 1996; and Roeder & Wasserman, 1997 for discussion of choice of priors for mixture model problems). This is in accordance with the empirical results obtained from the first simulation study, where unstable estimates resulted from specifying non-informative uniform priors for the slope and intercept parameters but not from specifying flat priors for the chance level parameter. Moreover, the simulation study indicates that even proper non-informative priors with $\sigma_\alpha^2 = 10^{10}$ or $\sigma_\beta^2 = 10^{10}$ do not help with convergence. For the 3PNO model, it is hence recommended that informative priors be adopted for the slope and intercept parameters to ensure convergence of the Markov chains and consequently to obtain valid and accurate posterior estimates. With respect to the chance level parameters, although the choice between non-informative vs. informative priors has relatively little impact on their posterior estimates, informative priors, such as $Beta(2, 7)$ or $Beta(5, 17)$, are suggested if the true parameters are close to 0 and a flat $Beta(1, 1)$ prior shall be adopted if the true parameters are away from 0.

For both 2PNO and 3PNO models, sample sizes play a more important role than test lengths, though it has to be pointed out that increased sample sizes tend to have a relatively larger effect on the posterior estimates of slope parameters than those of intercept parameters in the 3PNO model. The informative priors considered in this study result in relatively smaller bias and error when the actual parameters in the 2PNO model are close to 0 and when the actual parameters in the 3PNO models are within ± 1 standard deviations away from their means. Further, with these priors, the Gibbs sampler for the 3PNO model with $n = 1000$ performs even worse than the that for the 2PNO model with $n = 100$, indicating that the complexity of the 3PNO model may not be offset by increasing sample sizes and/or test lengths. In spite of that, similar conclusions are drawn with respect to the use of informative priors for both models. That is, when true item parameters are close to their prior means, prior distributions with small variances do improve estimation accuracy no matter how large the dataset is; and they have an added advantage of enhancing

computational efficiency when sample sizes are small, as considerably fewer iterations are needed for the Markov chains to converge.However, when they are away from their prior means, small prior variances are more likely to give rise to biased posterior estimates and hence should be avoided. As noted earlier, the results pertaining to the effect of the distance between true item parameters and prior means shed lights on situations where informative priors are correctly or incorrectly specified. Hence, these conclusions echo with what has been documented in the literature.

Given the fact that appropriate specifications can be sometimes tricky in the presence of various prior distributions, care has to be taken in specifying prior information for the item parameters in an IRT model. The choice of prior distributions has been an important consideration in Bayesian statistics (e.g., Robert, 2001). Often, the choice is between informative and non-informative priors. Using informative prior distributions allows the incorporation of information available to practitioners from the literature and in light of their experience with other test items. On the other hand, using non-informative priors implies that no information from the test theory is relevant to the parameter in question. For small datasets, the likelihood is not very informative relative to the prior information. Hence, the use of non-informative priors can lead to an extremely long computing time when using the Gibbs sampler to obtain posterior estimates. Alternatively, appropriately specified informative priors help improve computational efficiency and reduce uncertainty considerably. Consequently, carefully thought out informative priors, such as those in the form of judge's ratings adopted in Swaminathan et al. (2003), are recommended provided that one can afford the time and cost for the procedure.

With regard to the choice among a large class of prior distributions, it is desirable if one considers proper prior distributions that reflect vague prior beliefs but regularize the posterior distribution and make the Markov chain more stable (Johnson & Albert, 1999), such as the third specification for item slope or intercept parameters considered in the simulations. Appropriate specifications of the prior distributions require an intimate understanding of both the test instrument and the group examinees. In certain applications, prior information about specific values of the item parameters may be available from previous experiments, such as the actual testing situations where items are selected from an item bank or they are used more than once for longitudinal studies. If such prior information is available, it should be incorporated using a suitable prior distribution for the item parameters. For example, if there is an interval $(-B, B)$ in which most of the intercept parameters are expected to fall, one can represent this prior constraint by imposing a $N(0, \sigma_\beta^2)$ on β_j by matching $2\sigma_\beta$ to the upper limit of the target interval, B. In some applications, however, the prior information is not readily available. One useful approach is to implement the MML or MB estimation, which has been found to be robust when sample sizes are large ($n \geq 1000$), to obtain a baseline set of item parameter estimates so that the prior distribution can be specified accordingly (Baker, 1992). If one is not certain about the appropriateness of the priors adopted for the item parameters, prior variances are not suggested to be set too small in small samples. With large datasets, however, it

is preferable to test the sensitivity of MCMC estimates to non-informative and informative priors than to choose an arbitrary informative prior that perhaps markedly biases the results.

One needs to note that there are a large number of choices for prior distributions for the IRT model parameters. This paper only considers a small class of informative priors, namely, the conjugate prior where the hyperparameters are specified. It is possible to adopt non-conjugate priors with the use of more complicated MCMC algorithms. Prior information can also be specified in a hierarchical fashion so that the hyperparameters for the item parameters are unknown and have their own prior distributions (e.g., Swaminathan & Gifford, 1982, 1985, 1986). This kind of hierarchical modeling actually estimates the hyperparameters instead of specifying them and shall offer advantages when prior information regarding item parameters is not available. Along similar lines, a hierarchical normal prior for the person parameter, such as that in Lindley and Smith (1972), can be adopted. In addition, this paper assumes that the parameters of a given type, e.g., γ_j, in the IRT model are exchangeable (Swaminathan & Gifford, 1985) and are independently and identically distributed. It is possible that they are interrelated and hence a multivariate prior can be specified to model the dependency among item parameters. Alternatively, one can consider adopting a different prior distribution for each parameter if exchangeability is questionable. There is, unfortunately, little research investigating the effect of these prior specifications on the quality of MCMC estimates for IRT models. Hence, additional simulation studies are needed to provide a set of guidelines for the choice of such a wide class of prior distributions for the item parameters.

It has to be pointed out that the *RMSE* or *bias* criterion provides one way to evaluate the accuracy of parameter estimates. Other criteria can be considered, such as the one based on the item response function, which takes into consideration the interaction among parameters. This study, however, only employed the *RMSE* criterion because its focus was mainly on item parameter recovery using the Gibbs sampler with the two IRT models, and hence the discrepancy between the true and estimated parameters was of primary concern. Furthermore, the *RMSE* is said to be the most stringent criterion because "it looks at the error in estimation in each parameter separately without compensating for the other parameters" (Swaminathan et al., 2003, p.50).

Finally, it is noted that the analysis of variance approach for evaluating the impact of the different factors (see e.g. Harwell, Stone, Hsu, & Kirisci, 1996) was not adopted in this study due to the non-convergence problem resulted from using non-informative priors with 3PNO models. Future studies can use it as another summative method.

References

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics,* **17**, 251–269.

Albert, J. H., & Ghosh, M. (2000). Item response modeling. In D. K. Dey, S. K. Ghosh & B.

K. Mallick (Eds.), *Generalized linear models: A Bayesian perspective* (pp. 173–193). New York: Marcel Dekker.

Altman, M., Gill, J., & McDonald, M. P. (2004). *Numerical issues in statistical computing for the social scientist*. New Jersey: Wiley.

Baker, F. B. (1990). Some observations on the metric of BILOG results. *Applied Psychological Measurement*, **14**, 139–150.

Baker, F. B. (1992). *Item response theory: Parameter estimation techniques*. New York: Marcel Dekker.

Baker, F. B. (1998). An investigation of the item parameter recovery characteristics of a Gibbs sampling procedure. *Applied Psychological Measurement*, **22**, 163–169.

Béguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, **66**, 541–562.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis* (2nd ed.). New York: Springer.

Birnbaum, A. (1968). The Logistic Test Model. In F. Lord, and M. Novick (Eds.), *Statistical Theories of Mental Test Scores* (pp. 397–423). Reading, Mass: Addison-Wesley Publishing Co.

Carlin, B. P. (1996). Hierarchical longitudinal modeling. In W. R. Gilks, S. Richardson & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 303–319). London: Chapman & Hall/CRC.

Carlin, B. P., & Louis, T. A. (2000). *Bayes and empirical Bayes methods for data analysis* (2nd ed.). London: Chapman & Hall.

Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, **46**, 167–174.

Chib, S., & Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, **49**, 327–335.

Diebolt, J., Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society* (Series B), **56**, 363–375.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. New Jersey: Lawrence Erlbaum Associates, Inc.

Fox, J-P. (2007). Multilevel IRT modeling in practice with the package mlirt. *Journal of Statistical Software*, **20**, 1–16.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis*. Boca Raton: Chapman & Hall/CRC.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, **7**, 457–511.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Analysis and Machine Intelligence,* **6**, 721–741.

Gifford, J. A., & Swaminathan, H. (1990). Bias and the effect of priors in Bayesian estimation of parameters of item response models. *Applied Psychological Measurement*, **14**, 33–43.

Glas, C. W., & Meijer, R. R. (2003). A Bayesian approach to person fit analysis in item response theory models. *Applied Psychological Measurement*, **27**, 217–233.

Harwell, M., & Janosky, J. E. (1991). An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in BILOG. *Applied Psychological Measurement*, **15**, 279–291.

Harwell, M., Stone, C. A., Hsu, H., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement*, **20**, 101–126.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applica-

tions. *Biometrika*, **57**, 97–109.

Johnson, V. E., & Albert, J. H. (1999). *Ordinal data modeling.* New York: Springer-Verlag.

Lim, R. G., & Drasgow, F. (1990). Evaluation of two methods for estimating item response theory parameters when assessing differential item functioning. *Journal of Applied Psychology*, **75**, 164–174.

Lindley, D. V., & Smith, A. F. M. (1972). Bayesian estimates for the linear model. *Journal of the Royal Statistical Society*, **34**, 1–41.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillside, New Jersey: Lawrence Erlbaum Associates.

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

Mathworks, Inc. (2005). *MATLAB: The language of technical computing* [Computer software]. Natick, MA: Author.

Mengersen, K., & Robert, C. (1996). Testing for mixtures: A Bayesian entropy approach. In J. Bernardo, J. Berger, A. David and A. Smith (Eds.), *Bayesian statistics 5* (pp. 255–268). Oxford: Oxford University Press.

Metropolis, N., & Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, **44**, 335–341.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, **51**, 177–195

Nylander, J. A., Wilgenbusch, J. C., Warren, D. L., & Swofford, D. L. (2008). AWTY (Are we there yet?): A system for graphical exploration of MCMC convergence in Bayesian phylogenetics. *Bioinformatics*, **24**, 581–583.

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

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

Richardson, S., & Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society* (Series B), **59**, 731–792.

Robert, C. P. (2001). *The Bayesian choice.* New York: Springer.

Roeder, K., & Wasserman, I. (1997). Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, **92**, 894–902.

Sahu, S. K. (2002). Bayesian estimation and model choice in item response models. *Journal of Statistical Computation and Simulation*, **72**, 217–232.

Swaminathan, H., & Gifford, J. A. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics*, **7**, 175–191.

Swaminathan, H., & Gifford, J. A. (1983). Estimation of parameters in the three-parameter latent trait model. In D. Weiss (Ed.), *New horizons in testing* (pp.13–30). New York: Academic Press.

Swaminathan, H., & Gifford, J. A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika*, **50**, 175–191.

Swaminathan, H., & Gifford, J. A. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika*, **51**, 581–601.

Swaminathan, H., Hambleton, R. K., Sireci, S. G., Xing, D., & Rizavi, S. (2003). Small sample estimation in dichotomous item response models: Effect of priors based on judgmental information on the accuracy of item parameter estimates. *Applied Psychological Measurement*, **27**, 27–51.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distribution by data augmentation (with discussion). *Journal of the American Statistical Association*, **82**, 528–550.

Tsutakawa, R. K., & Johnson, J. C. (1990). The effect of uncertainty of item parameter estimation on ability estimates. *Psychometrika*, **55**, 371–390.

Tsutakawa, R. K., & Soltys, M. J. (1988). Approximation for Bayesian ability estimation. *Journal of Educational Statistics*, **13**, 117–130.

Wollack, J. A., Bolt, D. M., Cohen, A. S., & Lee, Y-S. (2002). Recovery of item parameters in the nominal response model: A comparison of marginal maximum likelihood estimation and Markov chain Monte Carlo estimation. *Applied Psychological Measurement*, **26**, 339–352.

Yen, W. M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika*, **52**, 275–291.

Zimowski, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG 3: Item analysis and test scoring with binary logistic models* [Computer software]. Chicago, IL: Scientific Software.

(Received February 3 2009, Revised April 20 2010)