DOCUMENT RESUME

ED 420 689 TM 028 364

AUTHOR Kim, Seock-Ho; Cohen, Allan S.
TITLE An Evaluation of a Markov Chain Monte Carlo Method for the Two-Parameter Logistic Model.
PUB DATE 1998-04-16
NOTE 41p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Bayesian Statistics; *Estimation (Mathematics); Higher Education; *Markov Processes; *Maximum Likelihood Statistics; *Monte Carlo Methods
IDENTIFIERS Accuracy; Gibbs Sampling; Law School Admission Test; *Two Parameter Model

ABSTRACT

The accuracy of the Markov Chain Monte Carlo (MCMC) procedure Gibbs sampling was considered for estimation of item parameters of the two-parameter logistic model. Data for the Law School Admission Test (LSAT) Section 6 were analyzed to illustrate the MCMC procedure. In addition, simulated data sets were analyzed using the MCMC, marginal Bayesian estimation, and marginal maximum likelihood estimation methods. Data were simulated with 100 or 300 examinees and 15 or 45 items. Two different priors, informative and uninformative, were employed in the MCMC procedure. Marginal Bayesian estimation yielded consistently smaller root mean square differences and mean Euclidean distances than the other estimation methods. An appendix provides additional information about the computations. (Contains 6 tables, 5 figures, and 52 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 420 689

An Evaluation of a Markov Chain Monte Carlo Method for the Two-Parameter Logistic Model

Seock-Ho Kim
The University of Georgia
Allan S. Cohen
University of Wisconsin–Madison

April 16, 1998
Running Head: MCMC METHOD FOR 2PL

Paper presented at the annual meeting of the American Educational Research Association, San Diego, California

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Seock-Ho Kim
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM028364

An Evaluation of a Markov Chain Monte Carlo Method for the Two-Parameter Logistic Model

Abstract

The accuracy of the Markov chain Monte Carlo (MCMC) procedure Gibbs sampling was considered for estimation of item parameters of the two-parameter logistic model. Data for the Law School Admission Test Section 6 were analyzed to illustrate the MCMC procedure. In addition, simulated data sets were analyzed using the MCMC, marginal Bayesian estimation, and marginal maximum likelihood estimation methods. Two different priors, informative and uninformative, were employed in the MCMC procedure. Marginal Bayesian estimation yielded consistently smaller root mean square differences and mean Euclidean distances than the other estimation methods.

Key words: Bayesian inference, Gibbs sampling, item response theory, Markov chain Monte Carlo, marginal maximum likelihood estimation, prior, posterior.

Introduction

Some problems in statistical inference require integration over possibly high-dimensional probability distributions in order to estimate model parameters of interest or to obtain characteristics of model parameters. One such problem is estimation of item and ability parameters in the context of item response theory (IRT). Except for certain rather simple problems with highly structured frameworks (e.g., an exponential family together with conjugate priors in Bayesian inference), the required integrations may not be analytically feasible. In this paper, we examine the accuracy of a set of strategies known as Markov Chain Monte Carlo (MCMC) methods for estimation of IRT item parameters. We focus on the accuracy of one particular MCMC procedure, Gibbs sampling (Geman & Geman, 1984), for estimation of item parameters for the two-parameter logistic (2PL) model.

A number of ways exist for implementing the MCMC method. [For a review, refer to Bernardo and Smith (1994), Carlin and Louis (1996), and Gelman, Carlin, Stern, and Rubin (1995).] Metropolis and Ulam (1949), Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (1953), and Hasting (1970) present a general framework within which the Gibbs sampling (Geman & Geman, 1984) can be considered as a special case. In this regard, Gelfand and Smith (1990) discuss several different Monte Carlo-based approaches, including Gibbs sampling, for calculating marginal densities. [See Gilks, Richardson, and Spiegelhalter (1996) for a recent survey of applications.] Basically Gibbs sampling is applicable for obtaining parameter estimates for the complicated joint posterior distribution in Bayesian estimation under IRT (e.g., Mislevy, 1986; Swaminathan & Gifford, 1985; Tsutakawa & Lin, 1986).

Albert (1992) applied Gibbs sampling in the context of IRT to estimate item parameters for the two-parameter normal ogive model and compared these estimates with those obtained using maximum likelihood estimation. Baker (in press) has also investigated item parameter recovery characteristics of Albert's Gibbs sampling method for item parameter estimation via a simulation study. Patz and Junker (1997) developed a MCMC method based on the Metropolis-Hasting algorithm and presented an illustration using the 2PL model.

MCMC computer programs in the context of IRT have been developed largely only for specific applications. For example, Albert (1992) used a computer program written in MATLAB (The MathWorks, Inc., 1996). Baker (in press) developed a specialized FORTRAN version of Albert's Gibbs sampling program to estimate item parameters of the two parameter normal ogive model. Patz and Junker (1997) developed an S-PLUS code

(MathSoft, Inc., 1995). Spiegelhalter, Thomas, Best, and Gilks (1997) have also developed a general Gibbs sampling computer program BUGS for Bayesian estimation, using the adaptive rejection sampling algorithm (Gilks & Wild, 1992). The computer program BUGS requires specification of the complete conditional distributions.

Marginal maximum likelihood estimation (MMLE) using the expectation and maximization (EM) algorithm, as implemented in the computer program BILOG (Mislevy & Bock, 1990), has become the standard estimation technique for obtaining item parameter estimates under IRT. The Gibbs sampling procedure approaches the estimation of item parameters using the joint posterior distribution rather than the marginal distribution. Even so, both methods should yield comparable item parameter estimates, when comparable priors are used and when ignorance or locally uniform priors are used. This paper was designed to evaluate this issue using the 2PL model. Specifically, estimation methods based on the two computer programs, BUGS and BILOG, were examined and compared.

Theoretical Framework

Marginalized Solutions

Consider binary responses to a test with n items by each of N examinees. A response of examinee i to item j is represented by a random variable Y_{ij}, where $i = 1(1)N$ and $j = 1(1)n$. The probability of a correct response of examinee i to item j is given by $P(Y_{ij} = 1|\theta_i, \xi_j) = P_{ij}$ and the probability of an incorrect response is given by $P(Y_{ij} = 0|\theta_i, \xi_j) = 1 - P_{ij} = Q_{ij}$, where θ_i is ability and ξ_j is the vector of item parameters.

For examinee i, there is an observed vector of dichotomously scored item responses of length n, $Y_i = (Y_{i1}, \ldots, Y_{in})'$. Under the assumption of conditional independence, the probability of Y_i given θ_i and the vector of all item parameters, $\xi = (\xi_1, \ldots, \xi_n)'$, is

$$p(Y_i|\theta_i, \xi) = \prod_{j=1}^{n} P_{ij}^{Y_{ij}} Q_{ij}^{1-Y_{ij}}. \tag{1}$$

The marginal probability of obtaining the response vector Y_i for examinee i sampled from a given population is

$$p(Y_i|\xi) = \int p(Y_i|\theta_i, \xi) p(\theta_i) d\theta_i, \tag{2}$$

where $p(\theta_i)$ is the population distribution of θ_i. Without loss of generality, we can assume that the θ_i are independent and identically distributed as standard normal, $\theta_i \sim N(0, 1)$. This assumption may be relaxed as the ability distribution can also be empirically characterized

(Bock & Aitkin, 1981). The marginal probability of Y_i can be approximated with any specified degree of precision by Gaussian quadrature formulas (Stroud & Secrest, 1966).

The marginal probability of obtaining the $N \times n$ response matrix Y is given by

$$p(Y|\xi) = \prod_{i=1}^{N} p(Y_i|\xi) = l(\xi|Y), \tag{3}$$

where $l(\xi|Y)$ can be regarded as a function of ξ given the data Y. In MMLE, the marginal likelihood is maximized to obtain maximum likelihood estimates of item parameters (see Bock & Aitkin, 1980).

Bayes' theorem tells us that the marginal posterior probability distribution for ξ given the data, Y, is proportional to the product of the marginal likelihood for ξ given Y and the prior distribution of ξ. That is,

$$p(\xi|Y) = \frac{p(Y|\xi)p(\xi)}{p(Y)} \propto l(\xi|Y)p(\xi), \tag{4}$$

where $\propto$ denotes proportionality. The marginal likelihood function represents the information obtained about ξ from the data. In this way, the data modify our prior knowledge of ξ. A prior distribution represents what is known about unknown parameters before the data are obtained. Prior knowledge or even relative ignorance can be represented by such a distribution. In marginal Bayesian estimation (MBE) of item parameters, the marginal posterior is maximized to obtain Bayes modal estimates of item parameters (see Mislevy, 1986).

Joint Estimation Procedures

Birnbaum (1968) and Lord (1980) describe the estimation of the θ and ξ by joint maximization of the likelihood function

$$p(Y|\theta,\xi) = \prod_{i=1}^{N}\prod_{j=1}^{n} P_j(\theta_i)^{Y_{ij}} Q_j(\theta_i)^{1-Y_{ij}} = l(\theta,\xi|Y), \tag{5}$$

where $\theta = (\theta_1, \ldots, \theta_N)'$. In implementation of joint maximum likelihood estimation (JMLE) (see Lord, 1986 for a comparison of marginalized and joint estimation methods), the item parameter estimation part for maximizing $l(\xi|Y,\hat{\theta})$ and the ability parameter estimation part for maximizing $l(\theta|Y,\hat{\xi})$ are iterated until a stable set of maximum likelihood estimates of item and ability parameters is obtained.

Extending the idea of joint maximization, Swaminathan and Gifford (1982, 1985, 1986) suggested that θ and ξ can be estimated by joint maximization with respect to the parameters of the posterior density

$$p(\theta, \xi|Y) = \frac{p(Y|\theta, \xi)p(\theta, \xi)}{p(Y)} \propto l(\theta, \xi|Y)p(\theta, \xi), \tag{6}$$

where $p(\theta, \xi)$ is the prior density of the parameters θ and ξ. This procedure is called joint Bayesian estimation (JBE). Under the assumption that priors of θ and ξ are independently distributed with probability density functions $p(\theta)$ and $p(\xi)$, the item parameter estimation part maximizing $l(\xi|Y, \hat{\theta})p(\xi)$, and the ability parameter estimation part maximizing $l(\theta|Y, \hat{\xi})p(\theta)$ are iterated to obtain stable Bayes modal estimates of item and ability parameters.

Gibbs Sampling

The main feature of MCMC methods is to obtain a sample of parameter values from the posterior density (Tanner, 1996). The sample of parameter values then can be used to estimate some functions or moments (e.g., mean and variance) of the posterior density of the parameter of interest. In comparison, in the above IRT estimation procedures via MMLE, MBE, JMLE, or JBE, the task is to obtain modes of the likelihood function or of the posterior distribution.

The Gibbs sampling algorithm is as follows (Gelfand & Smith, 1990; Tanner, 1996). First, instead of using θ and ξ, let ω be a vector of parameters with k elements. Suppose that the full or complete conditional distributions, $p(\omega_i|\omega_j, Y)$, where $i = 1(1)k$ and $j \neq i$, are available for sampling. That is, samples may be generated by some method given values of the appropriate conditioning random variables. Then given an arbitrary set of starting values, $\omega_1^{(0)}, \ldots, \omega_k^{(0)}$, the algorithm proceeds as follows:

Draw $\omega_1^{(1)}$ from $p(\omega_1|\omega_2^{(0)}, \ldots, \omega_k^{(0)}, Y)$,
Draw $\omega_2^{(1)}$ from $p(\omega_2|\omega_1^{(1)}, \omega_3^{(0)}, \ldots, \omega_k^{(0)}, Y)$,
$\vdots$
Draw $\omega_k^{(1)}$ from $p(\omega_k|\omega_1^{(1)}, \ldots, \omega_{k-1}^{(1)}, Y)$,
Draw $\omega_1^{(2)}$ from $p(\omega_1|\omega_2^{(1)}, \ldots, \omega_k^{(1)}, Y)$,
Draw $\omega_2^{(2)}$ from $p(\omega_2|\omega_1^{(2)}, \omega_3^{(1)}, \ldots, \omega_k^{(1)}, Y)$,
$\vdots$

Draw $\omega_k^{(2)}$ from $p(\omega_k|\omega_1^{(2)},\ldots,\omega_{k-1}^{(2)},Y)$,

$\vdots$

Draw $\omega_1^{(t+1)}$ from $p(\omega_1|\omega_2^{(t)},\ldots,\omega_k^{(t)},Y)$,

Draw $\omega_2^{(t+1)}$ from $p(\omega_2|\omega_1^{(t+1)},\omega_3^{(t)},\ldots,\omega_k^{(t)},Y)$,

$\vdots$

Draw $\omega_k^{(t+1)}$ from $p(\omega_k|\omega_1^{(t+1)},\ldots,\omega_{k-1}^{(t+1)},Y)$,

$\vdots$

The vectors $\omega^{(0)},\ldots,\omega^{(t)},\ldots$ are a realization of a Markov chain with a transition probability from $\omega^{(t)}$ to $\omega^{(t+1)}$ given by

$$p(\omega^{(t)},\omega^{(t+1)}) = \prod_{l=1}^{k} p(\omega_l^{(t+1)}|\omega_j^{(t)}, j>l, \omega_j^{(t+1)}, j<l, Y). \tag{7}$$

The joint distribution of $\omega^{(t)}$ converges geometrically to the posterior distribution $p(\omega|Y)$ as $t\to\infty$ (Geman & Geman, 1984, Bernardo & Smith, 1994). In particular, $\omega_i^{(t)}$ tends to be distributed as a random quantity whose density is $p(\omega_i|Y)$. Now suppose that there exist m replications of the t iterations. For large t, the replicates $\omega_{i1}^{(t)},\ldots,\omega_{im}^{(t)}$ are approximately a random sample from $p(\omega_i|Y)$. If we make m reasonably large, then an estimate, $\hat{p}(\omega_i|Y)$, can be obtained either as a kernel density estimate derived from the replicates or as

$$\hat{p}(\omega_i|Y) = \frac{1}{m}\sum_{l=1}^{m} p(\omega_i|\omega_{jl}^{(t)}, j\neq i, Y). \tag{8}$$

In the context of IRT, Gibbs sampling tries to obtain or sample sets of parameters from the joint posterior density $p(\theta,\xi|Y)$. Inferences with regard to parameters can then be made using the sampled parameters. Note that inference for both θ and ξ can be made from the Gibbs sampling procedure. In this paper, as in the marginalized solutions, we are particularly interested in the accuracy of the MCMC procedure for estimating item parameters.

An Example

Steps for Gibbs Sampling

The following example is presented using the familiar Law School Admission Test Section 6 (LSAT6) data from Bock and Lieberman (1970) (see also Bock & Aitkin, 1981). Model parameters were estimated by Gibbs sampling using the computer program BUGS

(Spiegelhalter et al., 1997). These same LSAT6 data have been analyzed under the Rasch model and under the two-parameter normal ogive (i.e., probit) model in Spiegelhalter, Thomas, Best, and Gilks (1996). Spiegelhalter, Thomas, et al. (1996) also compared the BUGS results with those from Bock and Aitkin (1981).

Gibbs sampling uses the following four basic steps (cf. Spiegelhalter, Best, et al., 1996):

1. Full conditional distributions and sampling methods for unobserved parameters must be specified.

2. Starting values must be provided.

3. Output must be monitored.

4. Summary statistics (e.g., estimates and standard errors) for quantities of interest must be calculated.

Discussion of the four steps involved are presented in detail below. In addition, comparisons with the results from MBE and MMLE as implemented in the computer program BILOG (Mislevy & Bock, 1990) are presented.

Model Specifications

The model specifications are used as input to the BUGS computer program. In the LSAT6 data set, the item responses Y_{ij} are independent, conditional on their parameters P_{ij}. For examinee i and item j, each P_{ij} is a function of the ability parameter θ_i, the item discrimination parameter a_j, and the item difficulty parameter b_j under the 2PL. The θ_i are assumed to be independently drawn from a standard normal distribution for scaling purposes. Figure 1 shows a directed acyclic graph (see Lauritzen, Dawid, Larsen, & Leimer, 1990; Whittaker, 1990; Spiegelhalter, Dawid, Lauritzen, & Cowell, 1993) based on these assumptions. (It can be noted that λ_j and ζ_j are used in Figure 1 instead of a_j and b_j.) The model can be seen as directed because each link between nodes is represented as an arrow. The model can also be seen as acyclic because it is impossible to return to a node after leaving. It is only possible to proceed by following the directions of the arrows. Each variable or quantity in the model appears as a node in the graph, and directed links correspond to direct dependencies as specified above. The solid arrow denotes the probabilistic dependency, while dashed arrows indicate functional or deterministic relationships. The rectangle designates observed data, and circles represent unknown quantities.

Insert Figure 1 about here

It may be helpful to use the following definitions: Let v be a node in the graph, and V be the set of all nodes. A parent of v is defined as any node with an arrow extending from it and pointing to v, and a descendant of v is defined as any node on a direct path beginning from v. For identifying parents and descendants, deterministic links should be combined so that, for example, the parent of Y_{ij} is P_{ij}. It is assumed in Figure 1 that, for any node v, if we know the value of its parents, then no other nodes would be informative concerning v except descendants of v.

Lauritzen et al. (1990) indicated that, in a full probability model, the directed acyclic graph model is equivalent to assuming that the joint distribution of all the random quantities is fully specified in terms of the conditional distribution of each node given its parents. That is,

$$P(V) = \prod_{v \in V} P(v|\text{parents}[v]), \tag{9}$$

where $P(\cdot)$ denotes a probability distribution. This factorization not only allows extremely complex models to be built up from local components, but also provides an efficient basis for the implementation of MCMC methods (Spiegelhalter, Best, et al., 1996).

Gibbs sampling via the BUGS computer program works by iteratively drawing samples from the full conditional distributions of unobserved nodes in Figure 1 using the adaptive rejection sampling algorithm (Gilks, 1996; Gilks & Wild, 1992). For any node v, the remaining nodes are denoted by $V - v$. It follows that the full conditional distribution, $P(v|V-v)$, has the form

$$\begin{aligned} P(v|V-v) &\propto P(v, V-v) \\ &\propto P(v|\text{parent}[v]) \prod_{w \in \text{children}[v]} P(w|\text{parents}[w]). \end{aligned} \tag{10}$$

The proportionality constant, which is a function of the remaining nodes, ensures that the distribution is a probability function that integrates to unity.

To analyze the LSAT6 data, we begin by specifying the forms of the parent and child relationships in Figure 1. Under the 2PL model, the probability that examinee i responds correctly to item j is assumed to follow a logistic function parameterized by the examinee's latent ability θ_i, the item discrimination parameter, a_j, and the item difficulty parameter, b_j.

For estimation purposes, we use the form $a_j(\theta_i - b_j) = \lambda_j\theta_i + \zeta_j$, where the slope parameter $\lambda_j = a_j$ and the intercept parameter $\zeta_j = -a_j b_j$. Hence,

$$P_{ij} = \frac{1}{1 + \exp[-a_j(\theta_i - b_j)]} = \frac{1}{1 + \exp[-(\lambda_j\theta_i + \zeta_j)]}. \tag{11}$$

Since Y_{ij} are Bernoulli with parameter P_{ij}, we can define

$$Y_{ij} \sim \text{Bernoulli}(P_{ij}) \tag{12}$$

and

$$\text{logit}(P_{ij}) = \lambda_j\theta_i + \zeta_j. \tag{13}$$

To complete the specification of a full probability model in for the BUGS computer program, prior distributions of the nodes without parents (i.e., θ_i, λ_j, and ζ_j) also need to be specified. We can define these priors in several different ways. We can impose priors on λ_j and ζ_j using a hierarchical Bayes approach (e.g., Swaminathan & Gifford, 1985; Kim, Cohen, Baker, Subkoviak, & Leonard, 1994). If it is preferred that the priors not be too influential, uninformative priors could be imposed. Alternatively, it may also be useful to include external information in the form of fairly informative prior distributions. According to Spiegelhalter, Best, et al. (1996), it is important to avoid causal use of standard improper priors in MCMC modeling, since these may result in improper posterior distributions.

Following Spiegelhalter, Thomas, et al. (1996), two prior distributions were chosen for the LSAT6 analyses: (1) $\lambda_j \sim N(0, 1)$ with $\lambda_j > 0$ and $\zeta_j \sim N(0, 100^2)$ and (2) $\lambda_j \sim N(0, 100^2)$ with $\lambda_j > 0$ and $\zeta_j \sim N(0, 100^2)$. An example input file for BUGS is given in the Appendix.

Starting Values

The choice of starting values (e.g., $\omega^{(0)}$) is not generally that critical as the Gibbs sampler (and most other MCMC algorithms as well) should be run long enough to be sufficiently updated from its initial states. It is useful, however, to perform a number of runs using different starting values to verify that the final results are not sensitive to the choice of starting values (Gelman, 1996). Raftery (1996) indicated that extreme starting values could lead to a very long burn-in or stabilization process.

In this example, three runs were performed using the LSAT6 data with three sets of starting values for λ_j and ζ_j, $j = 1(1)5$. The first run started at values considered plausible in the light of the usual range of item parameters ($\lambda_j = 1$ and $\zeta_j = 0$). The second run at

$\lambda_j = 10$ and $\zeta_j = 5$ and the third at $\lambda_j = .1$ and $\zeta_j = -5$ represented substantial deviations in initial values. In particular, the second run was intended to represent a situation in which there was a possibility that items were highly discriminating, and the third run represented an opposite assumption. The priors used were $\lambda_j \sim N(0, 1)$ with $\lambda_j > 0$ and $\zeta_j \sim N(0, 100^2)$.

The three runs consisted of 10,000 iterations. Partial results for λ_1 and ζ_1 are presented in Figure 2. The computer program CODA (Best, Cowles, & Vines, 1997) was used to obtain these graphs. The top two plots in Figure 2 contain the graphical summaries of the Gibbs sampler for λ_1. The top left plot shows the trace of the sampled values of λ_1 for the three runs. In the legend, 1N0 indicates the initial values for λ_j and ζ_j were 1 and 0, respectively; 10N0 indicates the initial values were 10 and 0, respectively; and, P1N-5 indicates initial values were .1 and -5, respectively. Results for all three runs show that the λ_1 generated by the Gibbs sampler quickly settled down regardless of the starting values. The top right graph shows the kernel density plot of the three pooled runs of 30,000 values for λ_1. The variability among the λ_1 values generated by the Gibbs sampler seems not to be too great. The distribution looks like a truncated normal form due to the positive constraints on λ_j.

Insert Figure 2 about here

The bottom two plots contain graphical summaries of the Gibbs sampler for ζ_1. The bottom left plot shows the trace of the sampled values of ζ_1 for all three runs. The ζ_1 generated by the Gibbs sampler quickly settled down regardless of the starting values. The bottom right graph shows the kernel density plot of the three pooled runs of 30,000 values for ζ_1. The variability of the λ_1 values seems not too great. The sampled values seem to be concentrated around 3, and the sample values seem to follow a normal distribution.

The results for other item parameter estimates were very similar to those from λ_1 and ζ_1. Overall, the starting values appear to were not have affected the final results. Useful starting values for IRT problems can be found from the noniterative minimum logit chi-square estimation solution (Baker, 1987) or from values based on Jensema (1976) and Urry (1974) as employed in BILOG. Use of "good" starting values, such as from the above methods, can avoid the time delay required by a lengthy burn in. Our experience with these starting values indicates $\lambda_j = 1$ and $\zeta_j = 0$ will work sufficiently well for applications under the 2PL. In subsequent analyses, therefore, the values, $\lambda_j = 1$ and $\zeta_j = 0$, were used as starting values.

Output Monitoring

A critical issue for MCMC methods is how to determine when one can safely stop sampling and use the results to estimate characteristics of the distributions of the parameters of interest. In this regard, the values for the unknown quantities generated by the Gibbs sampler can be graphically and statistically summarized to check mixing and convergence. The method proposed by Gelman and Rubin (1992) is one of the most popular for monitoring Gibbs sampling. [Cowles and Carlin (1996) presented a comparative review of convergence diagnostics for MCMC algorithms.]

We illustrate, here, the use of Gelman and Rubin (1992) statistics on two 10,000 iteration runs. Details of the Gelman and Rubin method are given by Gelman (1996). Each 10,000 iteration run required about 160 minutes on a Pentium 90 megahertz computer. Monitoring was done using the suite of S-functions called CODA (Best et al., 1997). Figure 3 shows the trace lines of the sampled values of λ_1 and ζ_1 for the two runs. The plots in Figure 3 indicate that the two runs settled down quickly. Gelman-Rubin statistics (i.e., shrink factors) are also plotted on the right side of Figure 3 for λ_1 and ζ_1, respectively. For both parameters, the medians were stabilized after about 3,000 iterations.

Insert Figure 3 about here

For each parameter, the Gelman-Rubin statistics estimate the reduction in the pooled estimate of variance if the runs were continued indefinitely. The Gelman-Rubin statistics should be near 1 in order to be reasonably assured that convergence has occurred. The median for λ_1 in the example was 1.01 and the 97.5 percentage point was 1.03. The median for ζ_1 was 1.00 and the 97.5 percentage point was 1.02. These values were very close to 1.0, indicating that reasonable convergence was realized for all parameters.

The Gelman-Rubin statistics can be calculated sequentially as the runs proceed, and plotted as in Figure 3. These plots as well as other plots for λ_j and ζ_j suggest the first 3,000 iterations of each run be discarded and the remaining samples be pooled. We used 5,000 iterations as burn-in and the subsequent 5,000 iterations for estimating.

BUGS and BILOG Parameter Estimates

The posterior mean of the Gibbs sampler was obtained for each item parameter. Two different sets of prior distributions were employed in the BUGS runs. The first set employed

an informative prior on $\lambda_j \sim N(0,1)$ and an uninformative prior on $\zeta_j \sim N(0,100^2)$. In addition, a constraint was imposed on the ranges of λ_j to allow only positive values (i.e., $\lambda_j > 0$). The prior distribution for λ_j limits possible values. MCMCI indicates this informative prior for λ_j where I indicates the prior is informative. The second set employed two uninformative prior distributions, $\lambda_j \sim N(0,100^2)$ with the constraint $\lambda_j > 0$ and $\zeta_j \sim N(0,100^2)$. This second set of priors is labeled MCMCU, where U indicates the priors are uninformative.

For BILOG runs, two procedures were used, MBE and MMLE. The default prior in BILOG for the 2PL is only on the item discrimination parameter as $p(\log a_j) = N(\mu_{\log a_j}, \sigma^2_{\log a_j}) = N(0,.5^2)$. Default options of BILOG yield MBE. For MMLE, no prior distributions were used.

Insert Table 1 about here

The information in Table 1 indicates that all estimation methods yielded similar results. Differences among estimates were relatively small, indicating the estimates from the four methods were comparable. MBE yielded smaller standard errors for all parameter estimates and MCMCU yielded somewhat larger standard errors.

Empirical Comparison

Simulation Conditions

Data were simulated under the following conditions: number of examinees ($N = 100, 300$), and number of items ($n = 15, 45$). The following estimation conditions were used: algorithm (MCMC, marginalized), and prior condition (informative, uninformative/none). The sample sizes and the test lengths were selected to emulate a situation in which estimation procedures and priors might have some impact upon item parameter estimates (e.g., Harwell & Janosky, 1991). Sample size and test length were completely crossed to yield four conditions.

Two estimation procedures were used, the MCMC method and the marginalized estimation method. For the MCMC method, an informative (MCMCI) and an uninformative (MCMCU) prior were used. For MCMCI, $\lambda_j \sim N(0,1)$ with the constraint $\lambda_j > 0$ and $\zeta_j \sim N(0,100^2)$. For MCMCU, $\lambda_j \sim N(0,100^2)$ with the constraint $\lambda_j > 0$ and $\zeta_j \sim N(0,100^2)$. For marginalized estimation via BILOG, two conditions were used, a prior on item discrimination (MBE) and no prior (MMLE).

Data Generation

The data sets used in this study were the same as those used in Kim et al. (1994). Item response vectors were generated via the computer program GENIRV (Baker, 1982) for the 2PL model. The generating parameters for item discrimination were distributed with mean 1.046 and variance .103 (i.e., standard deviation .321), and the underlying item difficulty parameters were distributed normal with mean 0 and variance 1. Item discrimination and item difficulty parameters for the 15-item test were set to have three different values (the number of items is given in parentheses): Item discrimination parameters were .66 (4), 1.0 (7), and 1.51 (4), and item difficulty parameters were -1.38 (4), .0 (7), and 1.38 (4). For the 45-item test, each of the item parameters was set to have five different values: Item discrimination parameters were .57 (4), .76 (9), 1 (19), 1.32 (9), and 1.77 (4), and item difficulty parameters were -1.9 (4), $-.95$ (9), 0 (19), .95 (9), and 1.9 (4). There was no correlation between item discrimination and difficulty parameters. The underlying ability parameters distribution was normal (0, 1) meaning that the underlying parameters were matched to the item difficulty distribution.

Four replications were generated for each of the sample size and test length conditions. Sixteen GENIRV runs were needed to obtain the data sets for the study.

Item Parameter Estimation

Each of the generated data sets was analyzed via the computer program BILOG (Mislevy & Bock, 1990) for MBE and MMLE, and via the computer program BUGS (Spiegelhalter et al., 1997) for MCMCI and MCMCU. In each estimation method, two prior conditions were employed. For example, the generated item response data set for the first replication of sample size 100 and test length 15 was analyzed by four computer runs.

For MBE, a lognormal prior on item discrimination with mean 0 and variance .25 [i.e., $\log a_j \sim N(0, .5^2)$] was used. This is the default prior specification in BILOG for estimation of item parameters of the 2PL model. MMLE, as implemented in BILOG, does not use any priors for item parameters. All defaults in BILOG for the 2PL model were used in the calibration except for the NOSlope option which was used for the MMLE condition.

Priors in the MCMCU condition for both item parameters, λ_j and ζ_j, were uninformative distributions. The prior distribution for λ_j was set to have a normal distribution with mean 0 and variance 10,000 [i.e., $\lambda_j \sim N(0, 100^2)$] with range restricted to yield positive values

of λ_j (i.e., $\lambda_j > 0$). The prior distribution for ζ_j was also set to have a normal distribution with mean 0 and variance 10,000. These prior distributions were similar to uninformative uniform distributions defined on the positive real number line for λ_j and on the entire real range for ζ_j.

In the MCMCI condition, an informative prior was used for λ_j but an uninformative prior was used for ζ_j. The prior distribution for λ_j was set to have a normal distribution with mean 0 and variance 1 [i.e., $\lambda_j \sim N(0,1)$] with range restricted to yield positive values of λ_j (i.e., $\lambda_j > 0$). The prior distribution for ζ_j was $\sim N(0, 100^2)$. The prior distribution for λ_j can be seen as a half normal distribution or the singly truncated normal distribution (Johnson, Kotz, & Balakrishnan, 1994). Since λ_j, without the range restriction, was sampled from a unit normal distribution, then $E(\lambda_j) = .798$ and $\text{Var}(\lambda_j) = .363$ (standard deviation .603). The prior distribution for ζ_j, however, was similar to the uniform distribution defined on the entire real line.

In the example, we assumed different prior distributions for item discrimination and difficulty via BUGS and BILOG, respectively. Consequently, the priors for MBE and MCMCI were not the same. Likewise, the specifications used in MMLE were not the same as the prior specifications employed in MCMCU.

Metric Transformation

In parameter recovery studies, such as the present one, comparisons between estimates and the underlying parameters require that the item parameter estimates obtained from different calibration runs be placed on a common metric with their underlying parameters (Baker & Al-Karni, 1991; Yen, 1987). Parameter estimation procedures under IRT yield metrics which are unique up to a linear transformation. To link both sets of estimates and parameters, it is necessary to determine the slope and intercept of the equating coefficients required for the transformation.

The estimates of the item parameters for each of the estimation procedures were placed on the scale of the true parameters before comparisons were made. The test characteristic curve method by Stocking and Lord (1983) as implemented in the computer program EQUATE (Baker, 1993) was used.

Evaluation Criteria

The empirical evaluation in this study involved four criteria: root mean square difference (RMSD), correlation, mean euclidean distance (MED), and bias. The RMSD is the square root of the average of the squared differences between estimated and true values. For item discrimination, for example, RMSD is $\left\{(1/n)\sum_{j=1}^{n}(\hat{a}_j - a_j)^2\right\}^{1/2}$.

Since it is possible that an estimation procedure may function better at recovery of one type of item parameter than at recovery of another, it is sometimes useful to consider a single index which can simultaneously describe the quality of recovery for both item parameters. MED, which is the average of the square roots of the sum of the squared differences between the discrimination and difficulty parameter estimates and their generating values, provides such an index (Rudin, 1976). MED is defined as $(1/n)\sum_{j=1}^{n}\left\{(\hat{\xi}_j - \xi_j)'(\hat{\xi}_j - \xi_j)\right\}^{1/2}$, where $\hat{\xi}_j = (\hat{a}_j, \hat{b}_j)'$ and $\xi_j = (a_j, b_j)'$. One caveat in using MED, of course, is that item discrimination and difficulty parameters are not expressed in comparable and interchangeable metrics. Even so, MED does provide a potentially useful descriptive index.

It is also useful to examine the bias, B, between the expected value of the estimates and the corresponding parameter. The bias of the item discrimination estimates, for example, is given as $B_{a_j} = E(\hat{a}_j) - a_j$. This estimate of bias was obtained for both parameters in the model across the four replications.

Results

RMSD and Correlation Results

Average RMSDs for item discriminations over four replications are reported in Table 2. As sample size increased, RMSDs decreased; marginal RMSD means were .328 and .228 for sample sizes 100 and 300, respectively.

Insert Table 2 about here

MBE consistently yielded the smallest RMDSs followed by MCMCI. For sample size 100, increasing the number of items decreased RMSD for MMLE but did not change RMSDs for the other cases. Increasing the number of items reduced the size of RMSDs for sample size 300. MCMCU tended to yield larger values of RMSD than the other estimation procedures in the 100-examinee by 15-item condition.

The average correlations between true and estimated values of item discriminations across four replications are also given in Table 2. Only very minor differences occurred between estimation methods. Generally, the larger the sample size and the longer the test, the higher the correlation, although these differences were small.

Table 3 contains the average RMSDs for item difficulty over four replications. An increase in sample size appears to be associated with a decrease in the size of RMSDs. RMSDs from MBE were consistently smaller than from the other estimation procedures. MCMCI yielded slightly smaller RMSDs than either MCMCU or MMLE.

Insert Table 3 about here

For each data set, all estimation procedures yielded very high correlations between estimates and underlying parameters (see Table 3). The larger sample size yielded slightly higher correlations. Increasing the number of items did not appear to affect the magnitude of the correlations for either sample size. MBE yielded consistently higher correlations than did other procedures.

MED Results

Average MEDs between item parameter estimates and underlying item parameters over four replications are reported in Table 4. MBE consistently yielded the smallest MEDs followed by MCMCI. MEDs decreased as the sample size increased, although this effect was quite small.

Insert Table 4 about here

Bias Results

The bias results for item discrimination, presented in both Table 5 and Figure 4, appear to reflect the influence of a number of factors. Each bias statistic was obtained by combining results from all four replications.

Insert Table 5 and Figure 4 about here

For the 15-item test, increasing sample size resulted in a decrease in bias values. When Bayes estimation procedures were used, it was expected that positive bias would be observed for the smaller item discrimination parameters (i.e., $a_j = .66$ for the 15-item test, and $a_j = .57$ and .76 for the 45-item test) and negative bias for the larger item discrimination parameters (i.e., $a_j = 1.51$ for the 15-item test, and $a_j = 1.32$ and 1.77 for the 45-item test). This shrinkage effect was observed only for MBE for sample size 100 and for MCMCI in the sample size 100 and the test length 15 condition.

MBE yielded different patterns of bias than did the other procedures. These differences were more evident for the 100-examinee condition, and diminished as the sample size increased 300. MCMCI yielded bias patterns somewhat closer to those for MBE. The patterns of bias for MCMCU and MMLE were similar.

The bias results for item difficulty are reported in Table 6 and Figure 5. The pattern of results was somewhat different from that for item discrimination. For the both 15- and 45-item tests, all estimation procedures yielded nearly the same pattern of essentially no bias. For the 100-examinee condition, MMLE yielded slightly larger bias. The 300-examinee condition yielded somewhat more stable bias results than did sample size 100.

Insert Table 6 and Figure 5 about here

Discussion

Previous work with the MCMC method using Gibbs sampling suggests this method may provide a useful alternative method for estimation when small sample sizes and small numbers of items are used. Even though implementation of the Gibbs sampling method in IRT is available in several computer programs, the accuracy of the resulting estimates have not been thoroughly studied.

The computer programs BUGS (Spiegelhalter et al., 1997) and CODA (Best et al., 1997) as well as the accompanying manuals are freely available over the Web. The uniform resource locator (URL) is:

`http://www.mrc-bsu.cam.ac.uk/bugs/`

The simulation results of this study indicate that MBE via BILOG yielded better item parameter estimates than other methods. A similar conclusion can be found in Baker (in press).

The Gibbs sampling and general MCMC methods are likely to be more useful for situations where complicated models are employed. For example, Gibbs sampling can be applicable to the estimation of item and ability parameters in the hierarchical Bayes approach (Mislevy, 1986; Swaminathan & Gifford, 1982, 1985, 1986). In this study the priors were imposed directly on the parameters. Accuracy of the Gibbs sampling method with different kinds of priors has not been investigated. This kind of research may be particularly valuable for small samples and short tests.

The focus in this paper was estimation of item parameters. One of the possible advantages of using Gibbs sampling or general MCMC methods, and something to consider in future research on these methods, is incorporation of uncertainly in item parameter estimates into estimation of ability parameters (cf. Patz & Junker, 1997).

References

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

Baker, F. B. (1982). *GENIRV: A program to generate item response vectors* [Computer program]. Madison, University of Wisconsin, Department of Educational Psychology, Laboratory of Experimental Design.

Baker, F. B. (1987). Item parameter estimation via minimum logit chi-square. *British Journal of Mathematical and Statistical Psychology, 40,* 50–60.

Baker, F. B. (1993). EQUATE 2.0: A computer program for the characteristic curve method of IRT equating. *Applied Psychological Measurement, 17,* 20.

Baker, F. B. (in press). An investigation of the item parameter recovery characteristics of a Gibbs sampling approach. *Applied Psychological Measurement.*

Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement, 28,* 147–162.

Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian theory.* Chichester, England: Wiley.

Best, N. G., Cowles, M. K., & Vines, S. K. (1997). CODA: Convergence diagnosis and output analysis software for Gibbs sampling output (Version 0.4) [Computer software]. Cambridge, UK: University of Cambridge, Institute of Public Health, Medical Research Council Biostatistics Unit.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 395–479). Reading, MA: Addison-Wesley.

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

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika, 35,* 179–197.

Carlin, B. P., & Louis, T. A. (1996). *Bayes and empirical Bayes methods for data analysis.* London: Chapman & Hall.

Cowles, M. K., & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association, 91,* 883–904.

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. (1996). Inference and monitoring convergence. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 131–143). London: Chapman & Hall.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis.* London: Chapman & Hall.

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 Transactions on Pattern Analysis and Machine Intelligence, 6,* 721–741.

Gilks, W. R. (1996). Full conditional distribution. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 75–88). London: Chapman & Hall.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov chain Monte Carlo in practice.* London: Chapman & Hall.

Gilks, W. R., & Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics, 41,* 337–348.

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

Hasting, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika, 57,* 97–109.

Jensema, C. (1976). A simple technique for estimating latent trait mental test parameters. *Educational and Psychological Measurement, 36,* 705–715.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1994). *Continuous univariate distributions* (2nd ed., Vol. 1). New York: Wiley.

Kim, S.-H., Cohen, A. S., Baker, F. B., Subkoviak, M. J., & Leonard, T. (1994). An investigation of hierarchical Bayes procedures in item response theory. *Psychometrika, 59,* 405–421.

Lauritzen, S. L., Dawid, A. P., Larsen, B. N., & Leimer, H.-G. (1990). Independence properties of directed Markov fields. *Networks, 20,* 491–505.

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

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement, 23,* 157–162.

MathSoft, Inc. (1995). S-PLUS (Version 3.3 for Windows) [Computer software]. Seattle, WA: Author.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics, 21,* 1087–1092.

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.

Mislevy, R. J., & Bock, R. D. (1990). BILOG 3: Item analysis and test scoring with binary logistic models [Computer software]. Mooresville, IN: Scientific Software.

Patz, R. J., & Junker, B. W. (1997). *A straightforward approach to Markov chain Monte Carlo methods for item response models* (Tech. Rep. No. 658). Pittsburgh, PA: Carnegie Mellon University, Department of Statistics.

Raftery, A. E. (1996). Hypothesis testing and model selection. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 163–187). London: Chapman & Hall.

Rudin, W. (1976). *Principles of mathematical analysis* (3rd ed.). New York: McGraw-Hill.

Spiegelhalter, D. J., Best, N. G., Gilks, W. R., & Inskip, H. (1996). Hepatitis B: a case study in MCMC methods. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 21–43). London: Chapman & Hall.

Spiegelhalter, D. J., Dawid, A. P., Lauritzen, S. L., & Cowell, R. G. (1993). Bayesian analysis in expert systems (with discussion). *Statistical Science, 8,* 219–283.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Gilks, W. R. (1996). *BUGS 0.5 examples* (Vol. 1, Version i). Cambridge, UK: University of Cambridge, Institute of Public Health, Medical Research Council Biostatistics Unit.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Gilks, W. R. (1997). BUGS: Bayesian inference using Gibbs sampling (Version 0.6) [Computer software]. Cambridge, UK: University of Cambridge, Institute of Public Health, Medical Research Council Biostatistics Unit.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201–210.

Stroud, A. H., & Secrest, D. (1966). *Gaussian quadrature formulas.* Englewood Cliff, NJ: Prentice-Hall.

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

Swaminathan, H., & Gifford, J. A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika, 50,* 349–364.

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

Tanner, M. A. (1996). *Tools for statistical inference: Methods for the exploration of posterior distributions and likelihood functions* (2nd ed.). New York: Springer-Verlag.

The MathWorks, Inc. (1996). MATLAB: The language of technical computing [Computer software]. Natick, MA: Author.

Tsutakawa, R. K., & Lin, H. Y. (1986). Bayesian estimation of item response curves. *Psychometrika, 51,* 251–267.

Urry, V. W. (1974). Approximations to item parameters of mental test models and their uses. *Educational and Psychological Measurement, 34,* 253–269.

Whittaker, J. (1990). *Graphical models in applied multivariate analysis.* Chichester: Wiley.

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

Table 1

LSAT6 Item Parameter Estimates and Standard Errors (s.e.)

		BUGS				BILOG			
		MCMCI		MCMCU		MBE		MMLE	
Parameter	Item	Estimate	(s.e.)	Estimate	(s.e.)	Estimate	(s.e.)	Estimate	(s.e.)
λ_j	1	.79	(.26)	.85	(.29)	.85	(.21)	.83	(.25)
	2	.72	(.19)	.73	(.23)	.76	(.16)	.72	(.19)
	3	.88	(.23)	.97	(.51)	.87	(.19)	.89	(.23)
	4	.69	(.19)	.66	(.19)	.74	(.16)	.69	(.19)
	5	.66	(.21)	.65	(.22)	.73	(.17)	.66	(.20)
ζ_j	1	2.78	(.21)	2.82	(.23)	2.79	(.18)	2.77	(.20)
	2	.99	(.09)	1.00	(.10)	1.00	(.09)	.99	(.09)
	3	.25	(.08)	.26	(.09)	.25	(.08)	.25	(.08)
	4	1.30	(.10)	1.28	(.10)	1.30	(.10)	1.30	(.10)
	5	2.07	(.14)	2.07	(.14)	2.09	(.13)	2.05	(.13)

Table 2

Root Mean Square Differences (RMSD) and Correlations for Item Discrimination Averaged Over Four Replications

			BUGS		BILOG	
	Sample	Item	MCMCI	MCMCU	MBE	MMLE
RMSD	100	15	.304	.372	.255	.412
	100	45	.302	.372	.255	.348
	300	15	.224	.271	.205	.254
	300	45	.199	.221	.181	.216
Correlation	100	15	.668	.668	.667	.657
	100	45	.677	.667	.679	.676
	300	15	.823	.811	.819	.815
	300	45	.864	.861	.863	.860

Table 3

Root Mean Square Differences (RMSD) and Correlation for Item Difficulty Averaged Over Four Replications

			BUGS		BILOG	
	Sample	Item	MCMCI	MCMCU	MBE	MMLE
RMSD	100	15	.329	.322	.315	.334
	100	45	.345	.355	.298	.352
	300	15	.200	.210	.174	.207
	300	45	.219	.223	.197	.224
Correlation	100	15	.951	.953	.955	.950
	100	45	.946	.944	.958	.942
	300	15	.983	.981	.987	.981
	300	45	.977	.976	.981	.975

Table 4

Mean Euclidean Distances (MED) Averaged Over Four Replications

		BUGS		BILOG	
Sample	Item	MCMCI	MCMCU	MBE	MMLE
100	15	.398	.436	.359	.451
100	45	.400	.440	.344	.423
300	15	.255	.277	.234	.268
300	45	.252	.266	.230	.262

Table 5

Bias Results for Item Discrimination

			BUGS		BILOG	
Sample	Item	Disc.	MCMCI	MCMCU	MBE	MMLE
100	15	.66	.160	.158	.19	.14
		1.00	−.001	.016	−.02	.00
		1.51	.044	.182	−.07	.23
100	45	.57	.104	.085	.16	.07
		.76	.052	.044	.08	.04
		1.00	.071	.095	.03	.08
		1.32	.050	.146	−.04	.12
		1.77	−.120	.037	−.25	−.01
300	15	.66	−.056	.047	.08	.05
		1.00	−.004	−.004	−.02	−.01
		1.51	.046	.141	−.02	.11
300	45	.57	−.036	−.042	.02	−.04
		.76	.066	.062	.07	.06
		1.00	.020	.024	.01	.02
		1.32	.002	.023	−.04	.01
		1.77	.143	.235	.07	.20

Table 6
Bias Results for Item Difficulty

			BUGS		BILOG	
Sample	Item	Diff.	MCMCI	MCMCU	MBE	MMLE
100	15	−1.38	.05	.05	.03	.07
		.00	−.00	−.01	.00	.00
		1.38	−.04	−.04	−.01	.05
100	45	−1.90	.03	.04	.05	.11
		−.95	−.08	−.10	−.06	−.08
		.00	.03	.03	.02	.03
		.95	.07	.04	.05	.09
		1.90	−.15	−.16	−.15	−.22
300	15	−1.38	−.02	−.03	−.01	−.01
		.00	.01	.01	.01	.01
		1.38	.06	.07	.04	.06
300	45	−1.90	.08	.07	.07	.11
		−.95	−.08	−.09	−.07	−.09
		.00	.03	.03	.03	.03
		.95	−.05	−.04	−.05	−.05
		1.90	.01	.01	.00	−.01

Figure Captions

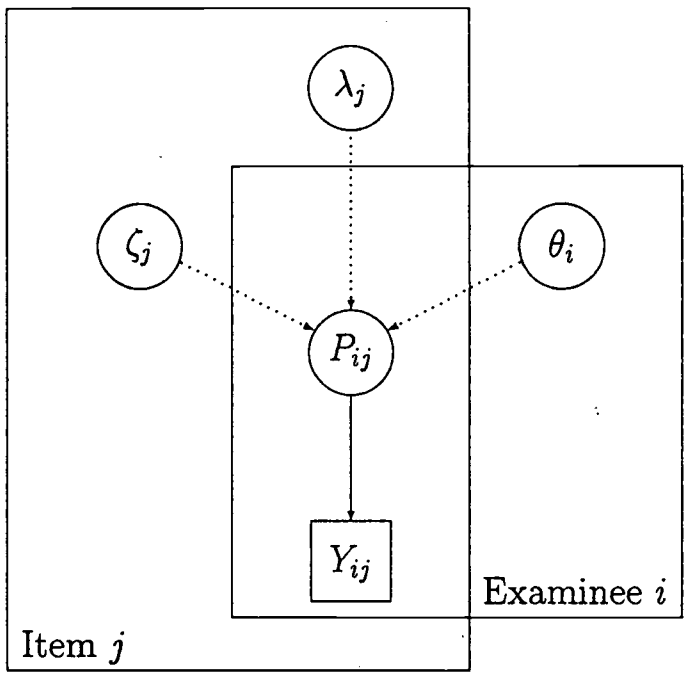
Figure 1. A directed acyclic graph for LSAT6 data.

Figure 2. Convergence with starting values for LSAT6 item 1 (λ_1 and ζ_1).

Figure 3. Traces plus Gelman and Rubin shrink factors for LSAT6 item 1 (λ_1 and ζ_1).

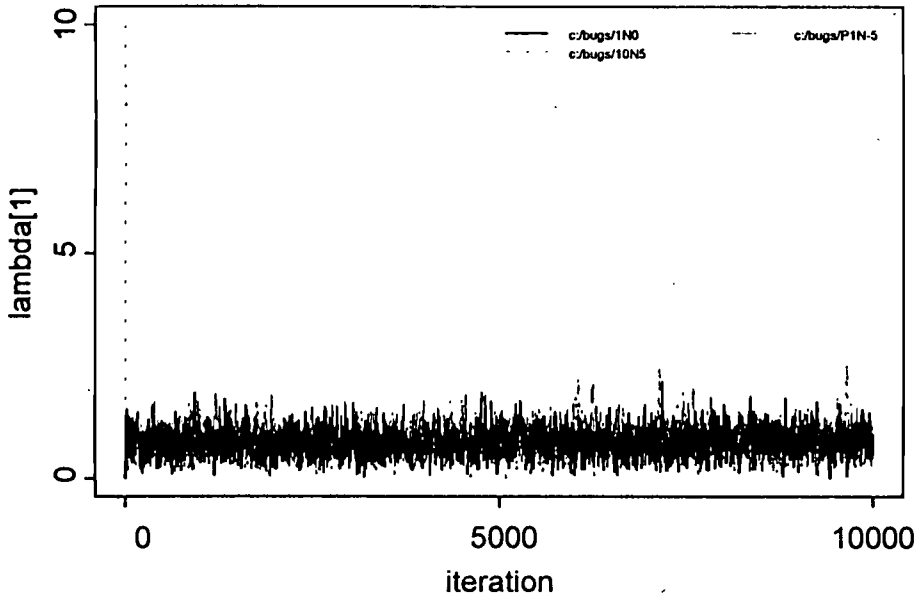
Figure 4. Bias plots for item discrimination.

Figure 5. Bias plots for item difficulty.

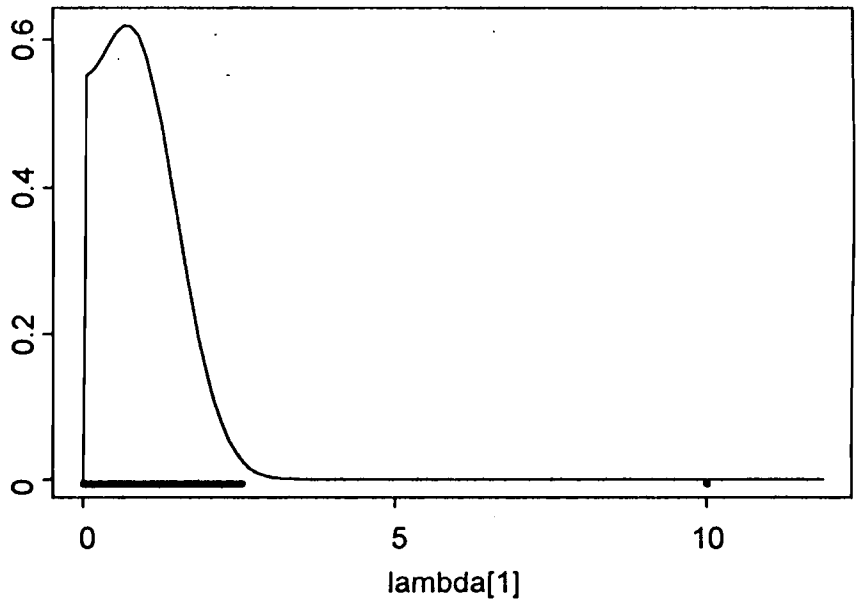

λ_j
ζ_j
θ_i
P_{ij}
Y_{ij}
Examinee i
Item j

Convergence with Starting Values for LSAT6 Item-1

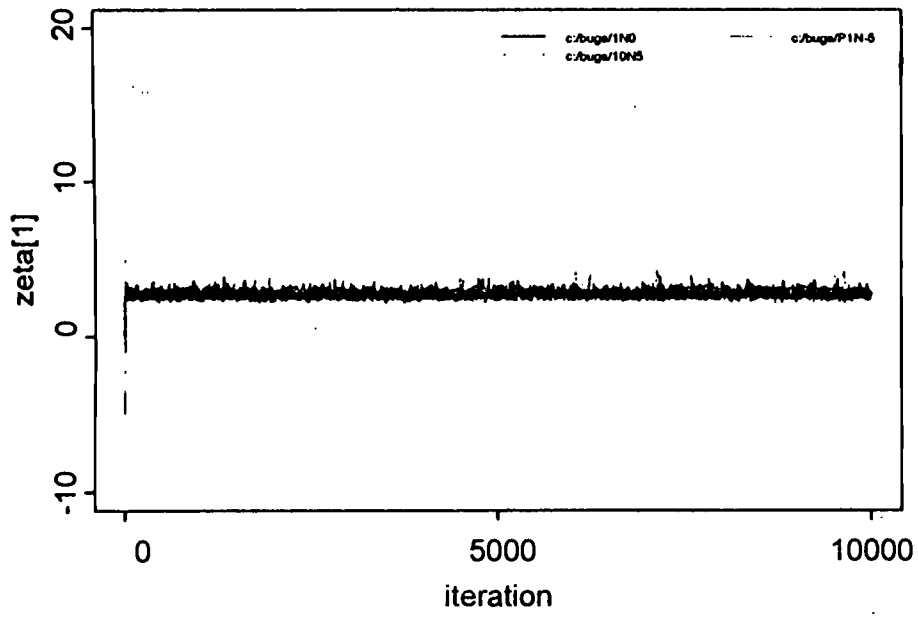
(10000 values per trace)



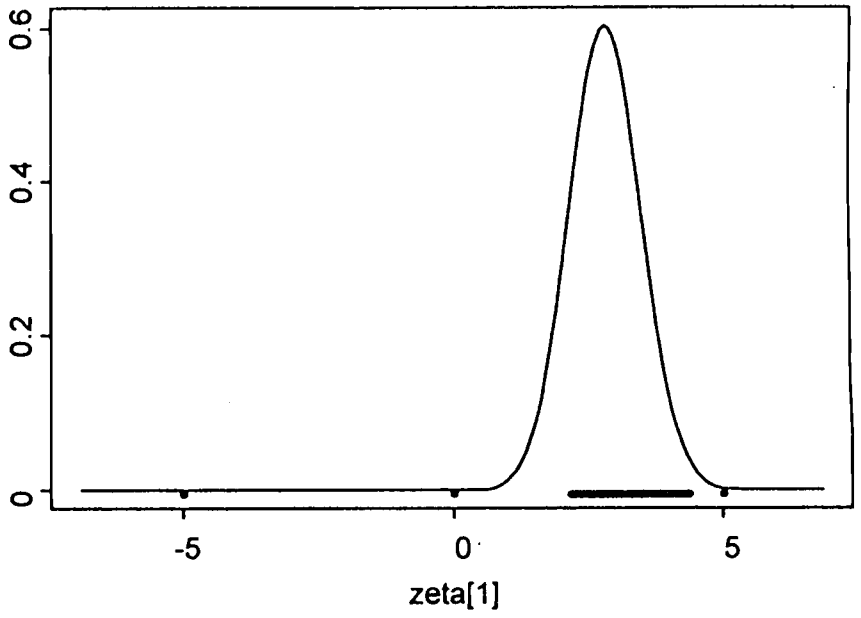
(30000 values)



(10000 values per trace)

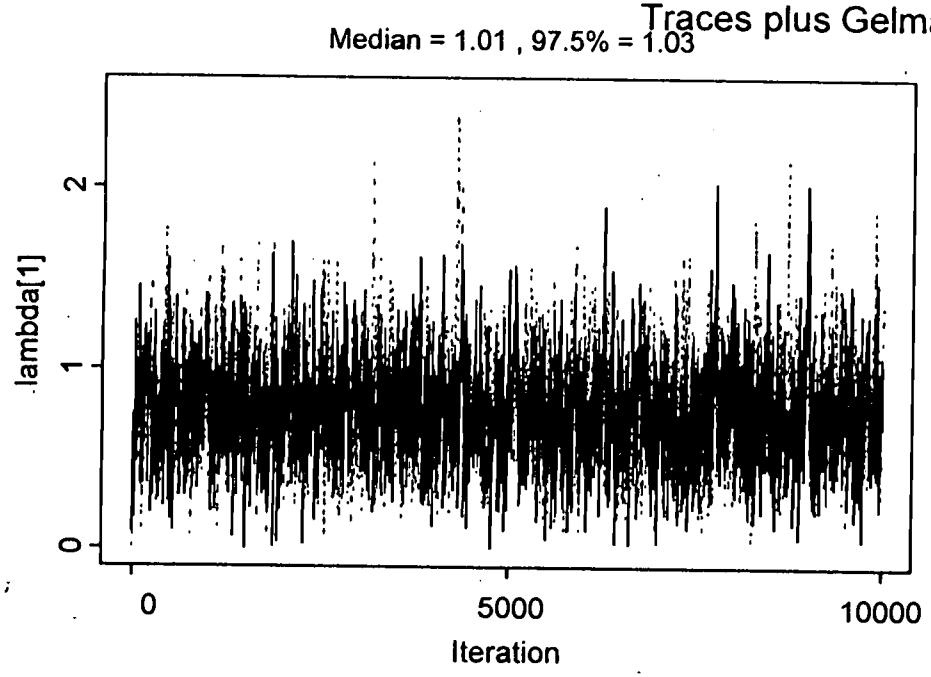


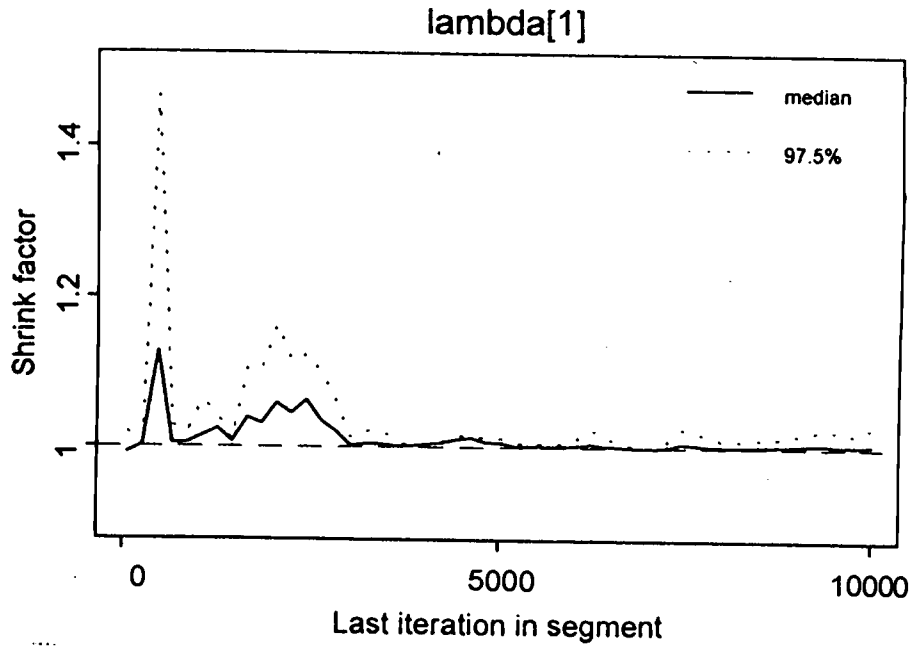
(30000 values)

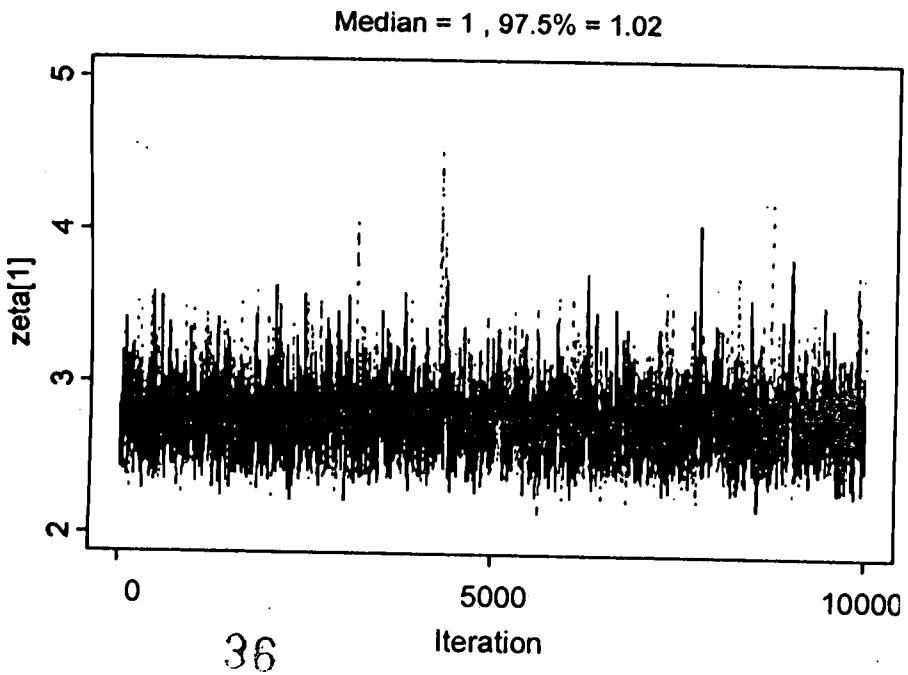


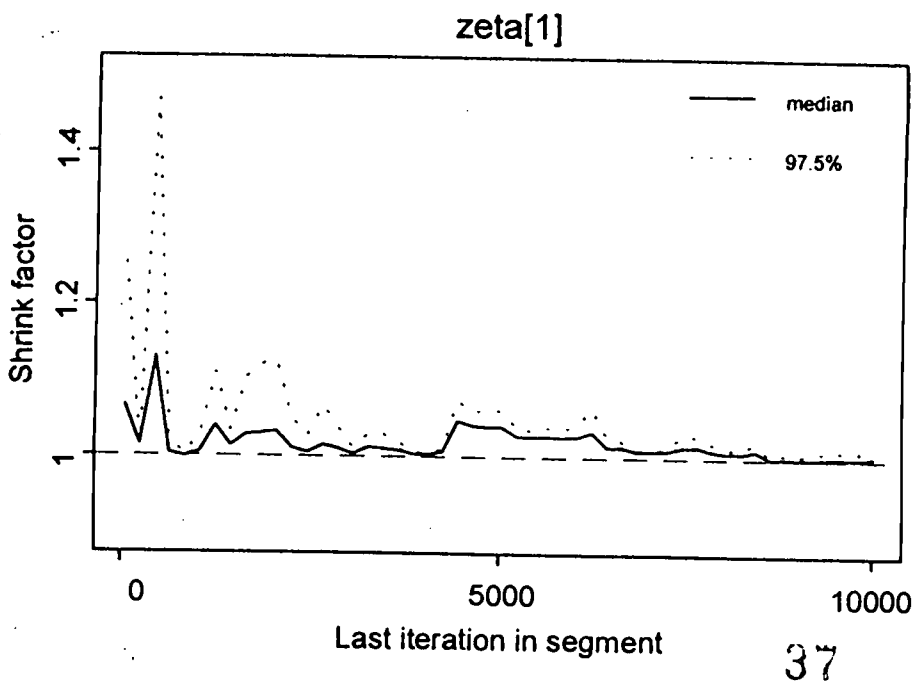
LSAT6 Item-1

Traces plus Gelman & Rubin Shrink Factors

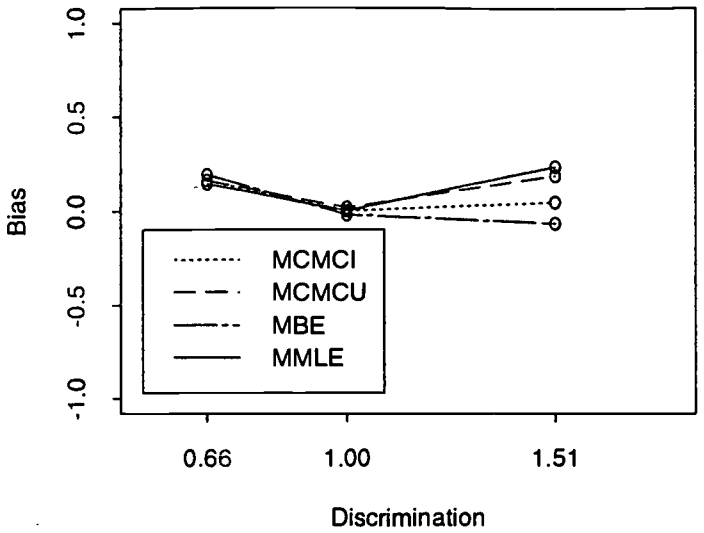




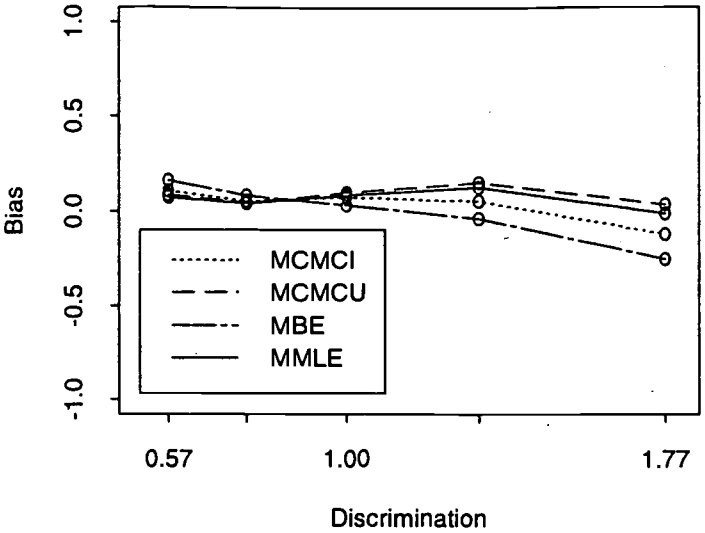




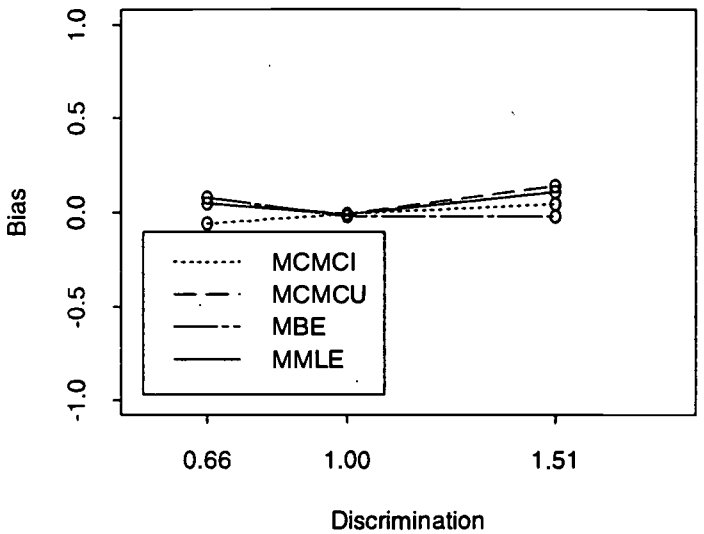
N=100, n=15



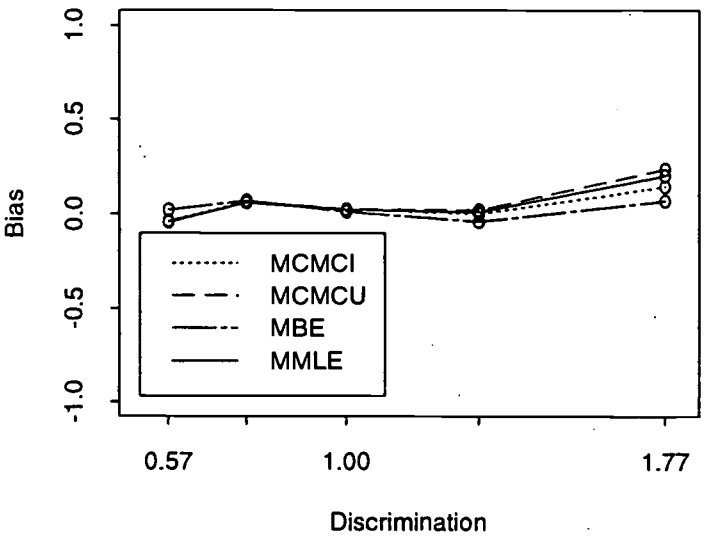
N=100, n=45



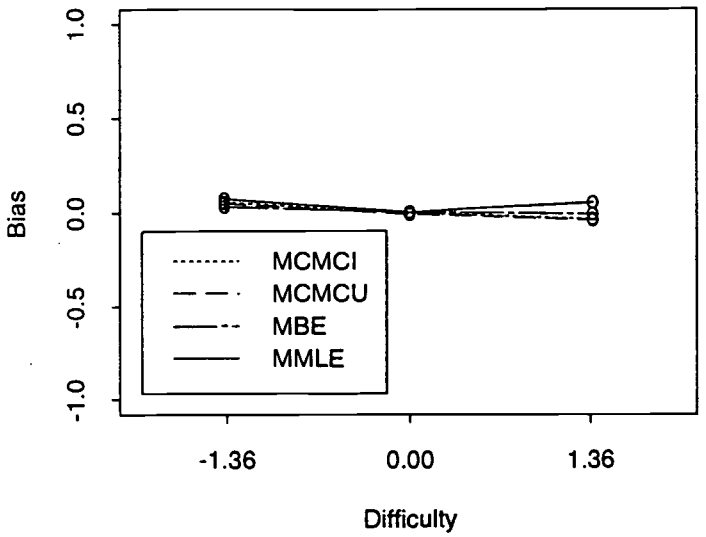
N=300, n=15



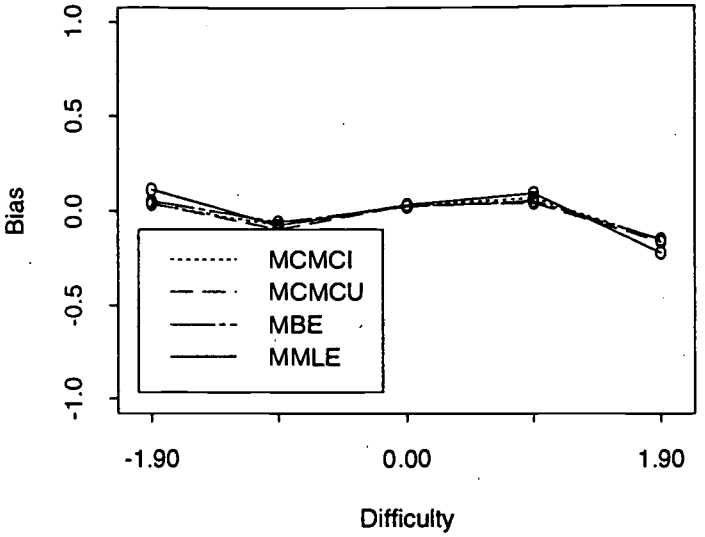
N=300, n=45



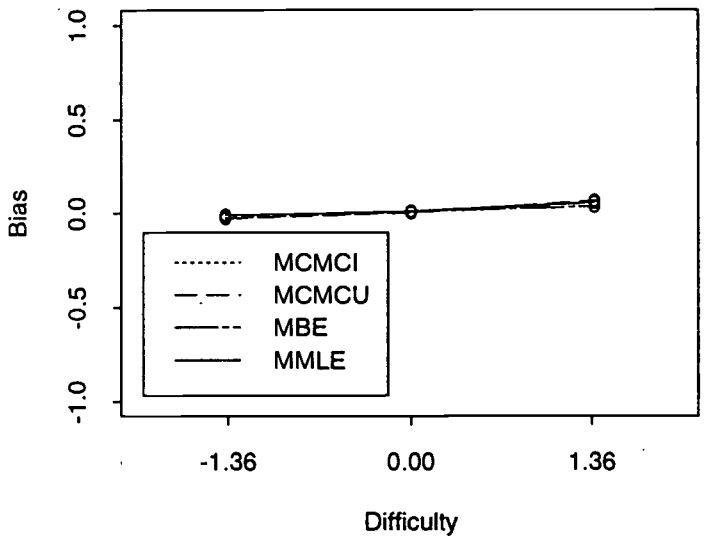
N=100, n=15



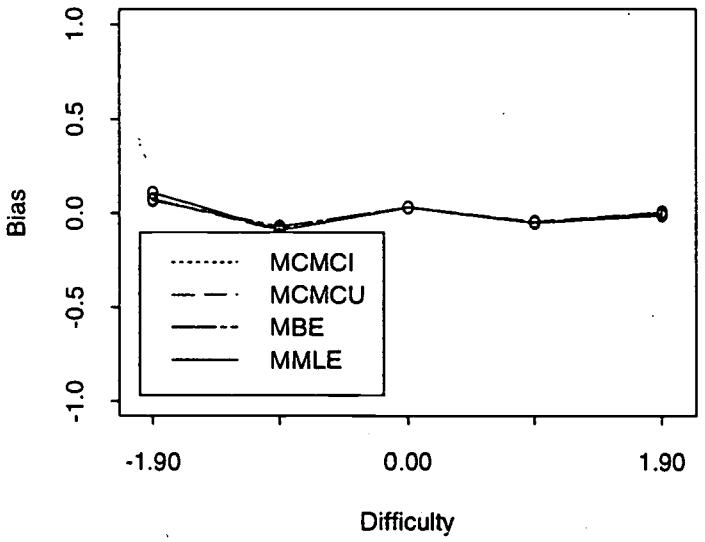
N=100, n=45



N=300, n=15



N=300, n=45



Appendix

```
model lsat6;
const
   I = 1000,
   J = 5;
var
   y[I,J], p[I,J], theta[I], lambda[J], zeta[J], b[J];
data in "lsat6-s.dat";
inits in "lsat6.in";
{
  for (i in 1:I) {
     for (j in 1:J) {
        logit(p[i,j]) <- lambda[j]*theta[i] + zeta[j];
        y[i,j] ~ dbern(p[i,j]);
     }
     theta[i] ~ dnorm(0,1);
  }
  for (j in 1:J) {
     lambda[j] ~ dnorm(0,1) I(0,);
     zeta[j] ~ dnorm(0,0.0001);
     b[j] <- - zeta[j]/lambda[j]
  }
}
```

Authors' Addresses

Send correspondence to Seock-Ho Kim, Department of Educational Psychology, The University of Georgia, 325 Aderhold Hall, Athens, GA 30602, Internet: skim@coe.uga.edu, or to Allan S. Cohen, Testing and Evaluation, 1025 West Johnson, Madison, WI 53706, Internet: ascohen@facstaff.wisc.edu.

TM028364

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: An Evaluation of a Markov Chain Monte Carlo Method for the Two-Parameter Logistic Model

Author(s): Seock-Ho Kim & Allan S. Cohen

Corporate Source: The University of Georgia
AERA

Publication Date: April, 1998

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 1

☑

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 2A

☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 2B

☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: Seock-Ho Kim

Printed Name/Position/Title: Seock-Ho Kim, Assistant Professor

Organization/Address: The University of Georgia
325 Aderhold Hall
Athens, GA 30602

Telephone: (706) 542-4224

FAX: (706) 542-4240

E-Mail Address: SKim@coe.uga.edu

Date: 4/3/98

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

The Catholic University of America
ERIC Clearinghouse on Assessment and Evaluation
210 O'Boyle Hall
Washington, DC 20064
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com

(Rev. 9/97)