Bayesian Estimation and Model Choice in Item Response Models

Sujit K. Sahu
Faculty of Mathematical Studies,
University of Southampton,
Highfield, Southampton, UK.
Email: S.K.Sahu@maths.soton.ac.uk

April 3, 2001

Abstract

Item response models are essential tools for analyzing results from many educational and psychological tests. Such models are used to quantify the probability of correct response as a function of unobserved examinee ability and other parameters explaining the difficulty and the discriminatory power of the questions in the test. Some of these models also incorporate a threshold parameter for the probability of the correct response to account for the effect of guessing the correct answer in multiple choice type tests.

In this article we consider fitting of such models using the Gibbs sampler. A data augmentation method to analyze a normal-ogive model incorporating a threshold guessing parameter is introduced and compared with a Metropolis-Hastings sampling method. The proposed method is an order of magnitude more efficient than the existing method. Another objective of this paper is to develop Bayesian model choice techniques for model discrimination. A predictive approach based on a variant of the Bayes factor is used and compared with another decision theoretic method which minimizes an expected loss function on the predictive space. A classical model choice technique based on a modified likelihood ratio test statistic is shown as one component of the second criterion. As a consequence the Bayesian methods proposed in this paper are contrasted with the classical approach based on the likelihood ratio test. Several examples are given to illustrate the methods.

Key words: Data Augmentation; Gibbs sampler; Markov chain Monte Carlo; Model Choice; Likelihood Ratio Statistic; Predictive Distribution; Three Parameter Model.

1 Introduction

Item response data are often found in educational and psychological testing. Typically a set of binary responses from a group of individuals on a number of questions constitute such data. Parametric item response models are popular for these data. The probability of correct response is modeled as a function of individual level effects due to the subjects and the questions (items). Models are built on additive and/or multiplicative assumptions on two types of parameters. To account for the effect of guessing in a multiple choice test another set of parameters, often called the guessing parameters, are also introduced, thus creating the so called three parameter models. These models appeared in item response theory (IRT) literature through the pioneering work of Birnbaum (1968), Lord (1980), Novick *et al.* (1972) and Rasch (1961) among many others. Accessible discussions and literature reviews are found in recent texts, e.g. Baker (1992), Johnson and Albert (1999), van der Linden and Hambleton (1997), and the references therein.

Model fitting in IRT is rather challenging since the joint maximum likelihood estimates (mle) for many of the models are not theoretically guaranteed to exist. Although some variants of the mle e.g. the conditional mle have finite values in some cases, usually they do not have the desirable large sample properties. This is because the number of unknown parameters under these models increases with the number of data points. General purpose software for these models, e.g. PC-BILOG (Mislevy and Bock, 1986), incorporate some limited Bayesian estimation techniques. Computational difficulties associated with a full Bayesian analysis are perhaps to blame for the paucity of literature in this area before the 1990s.

Computational techniques based on the Gibbs sampler (Gelfand and Smith, 1990) and other Markov chain Monte Carlo (MCMC) methods, see e.g. Gilks *et al.* (1996) for a review, have enabled routine fitting of these models under the Bayesian paradigm. In the IRT setting: Albert (1992) and Albert and Chib (1993) propose the Gibbs sampler for normal-ogive models; Ghosh *et al.* (2000) examine integrability of the associated posterior distributions; Patz and Junker (1999a, 1999b) describe and use more advanced MCMC methods. The Gibbs sampling software `BUGS` is also able to fit some of the models in this context (Spiegelhalter *et al.*, 1996).

The present article sets out with the following two main objectives.

- Introduction of an efficient data augmentation scheme (DAGS) for fitting the *normal-ogive models*.
- Comparison of different models using Bayesian predictive model selection criteria.

The new data augmentation scheme extends the work of Albert (1992) to fit the three parameter models and provides a much faster implementation of the Gibbs sampler. This is especially beneficial here since in many practical situations the available data sets are quite large and huge computing power is needed to fit these models. In one of our examples the proposed data augmentation scheme is approximately 10 times more efficient than a Metropolis-Hastings method currently used.

Another objective of this paper is to develop model selection techniques for deciding which model is the best fit for a given set of data. We propose doing this via a Bayesian model choice methodology and in so doing hope to address a range of modeling issues of which the most obvious include the following. Should we use a either logistic or a normal-ogive model? Is it worthwhile to include item discriminatory parameters? Should we include a set of guessing parameters? To our knowledge, the answers to these questions are not available in the IRT literature.

Model selection methods based on the classical likelihood ratio statistics (Andersen, 1973; Baker, 1992; Bock and Aitkin, 1981) and the Pearson χ^2 goodness of fit also run into problems for these models, because the asymptotics needed for the sample statistics are not fulfilled, see e.g. Lord (1975). Owing to these difficulties we consider Bayesian predictive model selection methods in this paper. A Bayesian formulation of the models does not rely on the asymptotic arguments and avoids many of the above mentioned difficulties, see e.g. Novick *et al.* (1972), Tsutakawa and Lin (1986), Swaminathan and Gifford (1986).

The methods based on features of the posterior distribution of the likelihood added to a penalty factor are available for model selection. For example, Aitkin (1997) interprets the p-values by using the posterior distribution of the likelihood function. Spiegelhalter *et al.* (1998) propose a model selection criterion for arbitrarily complex models called the deviance information criterion (DIC). They estimate the effective number of parameters for such models using quantities similar to the leverages in linear models. The penalty factor, which is the expected deviance minus the deviance evaluated at the posterior expectations, is calculated and added to the posterior expectation of the deviance to form the DIC.

In this paper we investigate the sensitivities of two model choice criteria available for Bayesian analysis. The first one, based on the pseudo-Bayes factor (Geisser and Eddy, 1979), is philosophically close to a pure Bayesian approach to model discrimination. The *Bayes factor* provides a measure of whether the observed data increased or decreased the relative odds of two different models under consideration.

The second approach considers a loss function based on the Kullback-Leibler type divergence measure between the observed set of data and a future replicate of the data arising under the fitted model. The expected loss with respect to the posterior *predictive distribution*, the distribution of a future replicate of the data, is the proposed criterion. From a given set of models, the one with the minimum expected loss is chosen as the best model for the data set currently under investigation. We show that the expected loss can be decomposed into two parts, one is the familiar likelihood ratio statistic and the other is a penalty for parameter estimation. In this way we are able to compare Bayesian model choice methods with conventional classical methods for model selection.

We also show that the above criteria can be used to distinguish between different sets of models based on different sets of assumptions. The proposed methods do not rely on asymptotic arguments and are by-products of the MCMC methods used to fit the models. In addition, the methods suggested here are based on predictive distributions and so do not require that the different models under investigation be nested, unlike the classical methods using the likelihood ratio tests.

The remainder of this article is organized as follows. Section 2 describes the models and the prior assumptions. In Section 3 we develop the DAGS and compare it with the Metropolis-Hastings methods in numerical examples. Section 4 describes the Bayesian model choice techniques with a large simulation example studying the sensitivities of different model choice criteria. We discuss a well known example in Section 5. We conclude with a few summary remarks in Section 6. Some computational details for calculating different model choice criteria are provided in the Appendix.

2 Hierarchical Models and Prior Assumptions

Suppose that each of n students are given k items (questions). The response y_{ij} for the ith student and the jth item is recorded as 1 or 0 according to whether the student answered the item correctly or incorrectly. Let p_{ij} denote the probability that the ith student is able to answer the jth item correctly.

We first consider the Rasch models (Rasch, 1961) where p_{ij} is modeled as

$$p_{ij} = F(\theta_i - \beta_j), \quad i = 1, \ldots, n, \ j = 1, \ldots, k, \tag{1}$$

where F is either the logistic cdf, i.e. $F(x) = \frac{1}{1+\exp(-x)}$ or the standard normal cdf $F(x) = \Phi(x)$. The models based on $\Phi(x)$ are known as normal-ogive models. The parameter θ_i measures the ability of the ith student and the parameter β_j measures the difficulty level of the jth item.

The two parameter models are obtained by introducing a slope parameter α_j (> 0) for each item. Here p_{ij} in (1) is modified to

$$p_{ij} = F(\theta_i \alpha_j - \beta_j), \quad \alpha_j > 0, i = 1, \ldots, n, \ j = 1, \ldots, k. \tag{2}$$

The restrictions $\alpha_j > 0, j = 1, \ldots, k$ assure that a student with a better ability θ_i has a higher probability of getting the jth item correct.

The three parameter models introduce a threshold probability c_j for p_{ij}. Here we assume

$$p_{ij} = c_j + (1 - c_j) F(\theta_i \alpha_j - \beta_j), \quad 0 \leq c_j < 1, \alpha_j > 0, i = 1, \ldots, n, \ j = 1, \ldots, k. \tag{3}$$

The differences between the three models are clear. The total probability of correct response under the three parameter model is explained by an additional factor of guessing. Contrasting the two and one parameter model, we can see that the two parameter model estimates a slope parameter α_j for each item while the one parameter model sets it at unity in an appropriate binary regression model. We denote the two and the three parameter models with the logistic cdf by `2PL` and `3PL` respectively (parameter logit). Similarly the normal-ogive models are denoted by `2PP` and `3PP` (parameter probit). The one parameter probit model is henceforth denoted by `1PP`.

Some authors prefer not to label the c_j, $j = 1, \ldots, k$ as guessing parameters. Instead those are viewed as threshold probabilities for p_{ij}. The quantity $\frac{1}{M}$ in a multiple choice type test with M alternatives serves as a very good guess for c_j. A beta prior distribution with density proportional to $x^{\kappa-1}(1-x)^{\lambda-1}$ for suitable non-negative values of κ and λ can be used for c_j. Higher values of these parameters lead to more precise prior information. These parameters should be chosen in such a way that $E(c_j) = \frac{\kappa}{\kappa+\lambda}$ is some pre-specified value, e.g. $\frac{1}{M}$, see e.g. Swaminathan and Gifford (1986). In many applications $M = 4$ and in this article we work with this choice throughout. Note that a uniform prior distribution gives $E(c_j) = \frac{1}{2}$.

The ability parameters $\theta_i, i = 1, \ldots, n$ are considered random effects following independent standard normal distributions, i.e. $\theta_i \sim N(0, \sigma^2)$ where $\sigma^2 = 1$. It is possible to estimate the ability variance σ^2 if it is unknown. In this case a suitable prior distribution for σ^2 is required. Conjugate inverse gamma prior distributions are usually considered, see e.g. Spiegelhalter *et al.* (1996). In this paper, however, we have not pursued this.

We assume a normal prior distribution, $N(0, \delta)$ say, for β_j and a truncated normal prior $N(0, \nu) I(\alpha_j > 0)$ where $I(\cdot)$ is the indicator function for α_j. Although we can fit the models with any suitable prior distributions, we choose the above for illustration. The truncated normal distribution prior for α_j, however, is advantageous to work with the normal-ogive models because of conjugacy. Also note that large values of δ and ν provide non-informative prior specification whereas smaller values correspond to higher prior precision.

Let $\boldsymbol{\zeta} = (\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \mathbf{c})$ denote the collection of parameters. The full joint posterior density, $\pi(\boldsymbol{\zeta}|\mathbf{y})$, of $\boldsymbol{\zeta}$ for the three parameter model is given by

$$\prod_{ij} \left[p_{ij}^{y_{ij}} (1 - p_{ij})^{1 - y_{ij}} \right] \exp \left[-\frac{1}{2} \left(\sum_{i=1}^{n} \theta_i^2 + \frac{1}{\delta} \sum_{j=1}^{k} \beta_j^2 + \frac{1}{\nu} \sum_{j=1}^{k} \alpha_j^2 \right) \right] \prod_j I(\alpha_j > 0) \prod_j [c_j^{\kappa - 1} (1 - c_j)^{\lambda - 1}]. \tag{4}$$

The posterior densities for the one and two parameter models are obtained appropriately. For example, the two parameter model is obtained by setting $c_j = 0$ and removing the prior term for c_j.

The models (2) and (3) are not identifiable since they are preserved if each θ_i is multiplied by a factor and each α_j is divided by the same, see e.g. Albert (1992). As claimed there, a proper prior for θ_i essentially gets rid of this difficulty. However, there may still be problems as α_j may be weakly identified by the full Bayesian model. This emerges from the fact that the precision assumed in the prior for θ_i together with the likelihood function may not be enough to 'separate' out θ_i from α_j. Hence the posterior distributions of some of the α_j may show evidence of very heavy tails. Here suitable choices of the hyper-parameters δ and ν can help. We illustrate the choices with two examples. Model parameters for the small data example in Section 3.3 with only 39 students and 6 items required high prior precision whereas the same for the second (larger) data set in Section 5 did not require such high precision.

3 Computation

3.1 MCMC using rejection and the Metropolis algorithm

The models described above can be fitted using the powerful MCMC methods such as the Gibbs sampler (Gelfand and Smith, 1990). For the item response models note that all the full conditional distributions are non-standard. Hence straightforward implementation of the Gibbs sampler using standard sampling distributions is not possible. However, all the full conditional distributions for the one and two parameter probit and logit models are *log-concave* (log of the density is concave), see e.g. Ghosh *et al.* (2000). Exact sampling from one dimensional log-concave distributions can be performed using rejection sampling, even when the normalizing constants are unknown, see e.g. Gilks and Wild (1992). These authors also develop an adaptive rejection sampling (ARS) scheme. ARS dynamically constructs two envelopes (one lower and one upper) for the distribution to be sampled from using successive evaluations of the density at the rejected points. The algorithm is stopped when one proposed point has been accepted. The routines to implement ARS are freely available from `http://www.mrc-bsu.cam.ac.uk/` and the Gibbs sampling software `BUGS` (also available from this web site) is able to fit these models.

Computational difficulty arises when fitting the three parameter models. The three parameter models are not contained within the broad framework of the one parameter exponential family models, rather they are mixtures of those. Hence log concavity of the full conditional distributions no longer holds. As a consequence nice computing tricks which are available for the exponential family models, e.g. ARS cannot be used here in general. However, the computations can be performed using a Metropolis step (Gilks *et al.*, 1995). Computer codes for performing adaptive rejection Metropolis sampling (ARMS) are publicly available from the above web site and we have used those for our examples in this paper.

However, MCMC samplers for the three parameter models with Metropolis sampling steps take a long time, e.g. days and hours, to run for large data sets, see e.g. Patz and Junker (1999a). To have faster implementations, we exploit the mixture structure in (3) to sample the guessing parameter c_j. Using standard techniques for mixture decomposition, we are able to obtain a sampling scheme in which all the required sampling can be performed using standard distributions for the `3PP` model, thus eliminating the need for Metropolis updating steps (see below).

3.2 Data Augmentation for the Normal-Ogive Models

We introduce two independent random variables corresponding to each data point y_{ij} as follows. The first is a Bernoulli random variable, denoted by u_{ij}, with success probability c_j. The second augmented variable is a normal random variable z_{ij} with mean $\eta_{ij} = \theta_i \alpha_j - \beta_j$ and variance unity. The response y_{ij} restricts the augmented random variables u_{ij} and z_{ij} so that the three parameter model (3) is obtained as a consequence. Thus we set

$$y_{ij} = u_{ij} + (1 - u_{ij}) \times I(z_{ij} > 0). \tag{5}$$

It is easy to see that (3) follows from (5) by taking expectations.

The Gibbs sampler corresponding to this data augmentation scheme is implemented as follows. First, we consider sampling u_{ij} and z_{ij} given the data and the other parameters. Note that the restriction (5) must be obeyed at every step of the algorithm. Consequently several cases arise depending on the value of the response y_{ij}. Suppose that $y_{ij} = 0$, then u_{ij} must be equal to zero and z_{ij} must be negative. Therefore, when $y_{ij} = 0$, we set $u_{ij} = 0$ and draw $z_{ij} \sim N(\eta_{ij}, 1)\, I(z_{ij} < 0)$.

Next we consider the case when $y_{ij} = 1$. If $u_{ij} = 0$ then z_{ij} must be positive, and if z_{ij} is negative then u_{ij} must be equal to unity. Hence we adopt the following conditional sampling scheme. If the current value of u_{ij} is 0 then we sample $z_{ij} \sim N(\eta_{ij}, 1)\, I(z_{ij} > 0)$. Otherwise z_{ij} is obtained as a draw from $N(\eta_{ij}, 1)$. Once z_{ij} has been sampled we test to see if it is negative. If it is, then we simply set $u_{ij} = 1$. Otherwise u_{ij} is drawn as a Bernoulli random variable with success probability c_j.

Now we draw a new value of c_j from the beta distribution with parameters $\kappa + \sum_{i=1}^n u_{ij}$ and $\lambda + n - \sum_{i=1}^n u_{ij}$. Once the u_{ij}, z_{ij} and c_j have been simulated we follow Albert's (1992) method for simulating the rest of the parameters, θ_i, α_j and β_j. Let $\tau = \sigma^{-2}$. The full conditional distribution for θ_i is normal with mean μ_b and variance V say, where

$$\mu_b = V^{-1}\omega \sum_{j=1}^k \alpha_j(z_{ij} + \beta_j), \text{ and } V^{-1} = \tau + \omega \text{ where } \omega^{-1} = \sum_{j=1}^k \alpha_j^2.$$

The full conditional distribution for the vector of parameters (α_j, β_j) is the bivariate normal distribution restricted over the range $\alpha_j > 0$ with mean $\boldsymbol{\mu}$ and dispersion Σ as given below,

$$\boldsymbol{\mu} = \Sigma^{-1} X^T Z_j, \text{ and } \Sigma^{-1} = X^T X + \mathrm{diag}(\delta^{-1}, \nu^{-1}),$$

where $\mathrm{diag}(\delta^{-1}, \nu^{-1})$ is a diagonal matrix; $X = [\mathbf{b}, -\mathbf{1}]$; $\mathbf{1}$ is an $n \times 1$ vector of 1's and $Z_j = [z_{1j}, \ldots, z_{nj}]^T$. Note the benefit of assuming conjugate normal priors for α_j and β_j. Non-conjugate priors except for few simple cases, e.g. an exponential distribution prior for α_j, will destroy the conjugate sampling distribution of α_j.

If the guessing parameter c_j needs to be fixed at some known value other than 0, we simply omit the sampling step for c_j. If however, they are not to be included at all, (as in the two and one parameter models) we set $u_{ij} = 0$ at every iteration and omit the sampling step for c_j.

It is straightforward to implement the above algorithm since all the complete conditional distributions for Gibbs sampling are standard. A computer program written in the C language is also available from the author.

3.3 An Example

An algorithm that is fast per iteration may produce highly autocorrelated samples, which are less useful for parameter estimation. In order to make fair comparison between different MCMC algorithms we use the notion of effective sample size, (ESS), Kass *et al.* (1989). ESS is defined for each parameter as the number of MCMC samples drawn, B, divided by the parameter's autocorrelation time, $\gamma = 1 + 2\sum_{k=1}^\infty \rho_k$, where ρ_k is the autocorrelation at lag k. Estimation of γ using sample autocorrelations is problematic because fewer MCMC samples are used in estimating ρ_k as k increases. There are many alternatives, see e.g. Roberts (1996) for a review. Here we use a simple upper bound $\frac{1+\rho^*}{1-\rho^*}$ where $\rho^* = \max_{k\geq 1} |\rho_k|$. In many applications $\rho^* = |\rho_1|$ and we have used this for our numerical example given below.

For the three parameter model in equation (4) the choices of the hyper-parameters, κ and λ play a significant role in the performance of the algorithms. As mentioned in Section 2, higher values of κ and

λ lead to more precise prior information. Hence the marginal densities of the parameters c_j become more peaked and concentrated around the prior mean $\kappa/(\kappa+\lambda)$.

Note that the DAGS samples the c_j parameters exactly from the conditional distribution. Although the Metropolis scheme also obtains exact samples, it is expected to be less efficient than the DAGS since effectively it has to search for the high density area in the unit interval from a very complicated conditional distribution obtained from Equation (4). Intuitively, the DAGS should out-perform the Metropolis scheme and the difference between the two should get better as more precise prior distributions are assumed. This point is illustrated empirically using the following example.

We consider a data set taken from Tanner (1996, page 190), see also Sahu and Roberts (1999) for an interesting study on the rate of convergence. For this example, we have $n = 39$ and $k = 6$. We have implemented the full three parameter probit model using both the DAGS and the ARMS. Although we have investigated many other choices for the hyper-parameters δ and ν, we take $\delta = 2$ and $\nu = \frac{1}{2}$ as in Patz and Junker (1999a). Larger values of these parameters led to unstable estimates.

We work with run lengths of 5000 iterations after discarding 1000 initial iterations. We obtain the effective sample size, ESS using the methodology described above for each of the 18 parameters (6 each of β_j, α_j and c_j). Also the sample size per second (ES/s), ESS divided by the running time, is calculated for each parameter. In Table 1 we report the average ESS and ES/s (average over 18 parameters) for the two algorithms run under different choices of the hyper-parameters κ and λ. The last column gives the ratio of the ES/s for the two algorithms.

	DAGS		ARMS		
	ESS	ES/s	ESS	ES/s	Eff
$\kappa = 1, \lambda = 1$	332.3	47.5	1377.3	5.6	8.9
$\kappa = 1, \lambda = 3$	341.2	48.7	1278.4	5.3	9.3
$\kappa = 6.25, \lambda = 18.75$	709.1	101.3	2011.5	8.1	11.5
$\kappa = 12.5, \lambda = 37.5$	962.6	137.5	2379.4	9.4	12.6
$\kappa = 100, \lambda = 300$	1642.0	234.6	3082.3	11.8	15.7

Table 1: Performance of the DAGS and ARMS for the Tanner's data example

We have included the uniform prior distribution for c_j in our analysis, although it does not give $E(c_j) = \frac{1}{2}$ a-priori. For this choice we see that the efficiency of the DAGS compared to the ARMS is minimum. All other choices have $E(c_j) = \frac{1}{4}$ a-priori. Note that, in terms of ESS the ARMS scheme does much better than the DAGS. However, when computing time is taken into account, the DAGS performs much better (about 10 times better). Lastly, the efficiency of the DAGS compared to the ARMS gets better as more precise prior information is assumed.

The performance of the DAGS gets even better for examples with larger data sets. For instance for the law school data example considered in Section 5 the DAGS is about 20 times more efficient than the ARMS. Note that these conclusions may change if one adopted different efficiency criteria from those considered here. However, we speculate that the performance of the sampling scheme under any sensible criteria would show an order of magnitude improvement similar to that reported here.

4 Model Choice

4.1 The Pseudo-Bayes Factor

A pure Bayesian approach to model selection is to report posterior probabilities of each model by comparing Bayes factors defined as follows. Let $\pi(\cdot)$ denote the density of its argument and $\boldsymbol{\zeta}$ denote all the parameters under the assumed model. Let $\mathbf{y}_{\text{obs}}$ denote the observed data with individual data points $y_{r,\text{obs}}, r = 1, \ldots, N$. The *prior predictive density* of a set of observations at the actual observed point $\mathbf{y}_{\text{obs}}$ is given by

$$\pi(\mathbf{y}_{\text{obs}}) = \int \pi(\mathbf{y}_{\text{obs}}|\boldsymbol{\zeta})\ \pi(\boldsymbol{\zeta})\ d\boldsymbol{\zeta}. \tag{6}$$

(Note that in the Bayesian inference setup the actual observations $\mathbf{y}_{\text{obs}}$ is fixed. The above is interpreted as the density of a set of observables evaluated at the observed point $\mathbf{y}_{\text{obs}}$.) The *Bayes factor* for comparing two given models M_1 and M_2 is

$$\text{BF} = \frac{\pi(\mathbf{y}_{\text{obs}}|M_1)}{\pi(\mathbf{y}_{\text{obs}}|M_2)},$$

where $\pi(\mathbf{y}_{\text{obs}}|M_i)$ is the density in (6) when M_i is the assumed model, $i = 1, 2$.

The BF gives a summary of the evidence for M_1 against M_2 provided by the data. Calibration tables for the BF are available for deciding how strong is the evidence, see e.g. Raftery (1996). Note that $\pi(\mathbf{y}_{\text{obs}}|M_i)$ is the marginal likelihood of the data under model M_i. Hence the BF chooses a model for which the marginal likelihood of the data is maximum.

Although there have been recent advances in computing the Bayes factor, see e.g. Raftery (1996) for a review, there are problems in calculating it for high dimensional models such as those advocated here. Also for improper priors the Bayes factor is not meaningful since it cannot be calibrated. This is because the predictive density (6) is improper when $\pi(\boldsymbol{\zeta})$ is. However, we can work with the following version of the BF which avoids these problems.

The methodology requires calculation of the cross-validation predictive densities. Let $\mathbf{y}_{(\text{r}),\text{obs}}$ denote the set of observations $\mathbf{y}_{\text{obs}}$ with rth component deleted. The *cross-validation predictive density* is defined by:

$$\pi(y_r|\mathbf{y}_{(\text{r}),\text{obs}}) = \int \pi(y_r|\boldsymbol{\zeta}, \mathbf{y}_{(\text{r}),\text{obs}})\ \pi(\boldsymbol{\zeta}|\mathbf{y}_{(\text{r}),\text{obs}})\ d\boldsymbol{\zeta}. \tag{7}$$

Note that in the case of conditionally independent observations given $\boldsymbol{\zeta}$, $\pi(y_r|\boldsymbol{\zeta}, \mathbf{y}_{(\text{r}),\text{obs}}) = \pi(y_r|\boldsymbol{\zeta})$. The above predictive density is also known as the conditional predictive ordinate (CPO). The *pseudo-Bayes factor* (PsBF) (Geisser and Eddy, 1979) for comparing two models M_1 and M_2 is defined as,

$$\text{PsBF} = \prod_{r=1}^{N} \frac{\pi(y_{r,\text{obs}}|\mathbf{y}_{(\text{r}),\text{obs}}, M_1)}{\pi(y_{r,\text{obs}}|\mathbf{y}_{(\text{r}),\text{obs}}, M_2)}.$$

This is a surrogate for the Bayes factor, see e.g. Gelfand (1996) and its interpretations are similar. The CPOs are also useful for checking model adequacy. Instead of using a single summary measure alone, e.g. the PsBF, the individual CPOs can also be compared under any two models. This is to guard against any single highly influential observation concealing a general trend. One observation, $y_{r,\text{obs}}$, prefers model M_1 to M_2 if the rth CPO is higher under M_1. The appendix contains details for calculating the PsBF.

4.2 Expected Predictive Deviance

For many Bayesian purists no further analysis is required after calculating the BF. In this article, however, we do not take such a strong view. Here we develop an alternative model selection criterion which provides an independent check for the conclusions obtained using the BF or the PsBF. Further, we shall see that the likelihood ratio statistic is one component of the criterion. This provides a way of comparing the Bayesian methods with the classical likelihood ratio tests.

The method is developed for a different but probabilistically equivalent description of the data and the models introduced in Section 2. We first describe this new setup. Recall that there are k items in our setting. Let $s = 2^k$ be the total number of response patterns. Hence the score vector of each of n students is a particular pattern among the total s possible patterns. Let the observed number of students with the rth pattern be denoted by $y_{r,\text{obs}}$. The resulting s dimensional count vector $\mathbf{y}_{\text{obs}}$ then follows a multinomial distribution with parameters n and an unknown probability vector $\mathbf{P}$ since the scoring pattern of each subject can be one and only one of the s patterns. In this setup we take $N = s$. Observe that $\mathbf{P}$ depends on the assumed model and the unknown parameters $\boldsymbol{\zeta}$, and its elements sum to unity. Later we shall see how to compute $\mathbf{P}$ using the assumed model.

Let $\mathbf{y}_{\text{rep}}$ (abbreviation for replicate) with components $y_{r,\text{rep}}, r = 1, \ldots, s$ denote a future set of observables under the assumed model. Intuitively, the assumed model is a 'good' fit to the observed data $\mathbf{y}_{\text{obs}}$ if $\mathbf{y}_{\text{rep}}$ is able to replicate the data well. Hence many model choice criteria can be developed by considering different loss functions measuring the divergence between $\mathbf{y}_{\text{obs}}$ and $\mathbf{y}_{\text{rep}}$. In particular, we consider the following loss function between the two

$$L(\mathbf{y}_{\text{rep}},\ \mathbf{y}_{\text{obs}}) = 2\left(\sum_{r=1}^{s} y_{r,\text{obs}} \log \frac{y_{r,\text{obs}}}{y_{r,\text{rep}}}\right). \tag{8}$$

Since (8) is an entropy like (Kullback-Leibler) divergence measure between $\mathbf{y}_{\text{obs}}/n$ and $\mathbf{y}_{\text{rep}}/n$, it is likely to yield high values if the predicted data $\mathbf{y}_{\text{rep}}$ is not 'close' to the observed data $\mathbf{y}_{\text{obs}}$. Furthermore, the rth term in the summation is strictly convex in $y_{r,\text{rep}}$ if $y_{r,\text{obs}}$ is positive. We can avoid such difficulties with the zero counts by removing the corresponding terms from the sum in (8) or add $\frac{1}{2}$ to every cell as is often done in practice, see e.g. Waller *et al.* (1997).

The best model among a given set of models is the model for which the expected value of the above loss function is the minimum, where the expectation is to be taken with respect to a suitable predictive distribution of $\mathbf{y}_{\text{rep}}$. Here the previously defined distributions, namely the prior predictive distribution (6) or the cross-validation distribution (7) can be considered. However, those have several limitations and are difficult to work with even in an MCMC setup. We consider the following posterior predictive density. This is similarly defined as (6) but the prior $\pi(\boldsymbol{\zeta})$ inside the integral is replaced by the posterior $\pi(\boldsymbol{\zeta}|\mathbf{y}_{\text{obs}})$. The *posterior predictive density* of $\mathbf{y}_{\text{rep}}$, given by

$$\pi(\mathbf{y}_{\text{rep}}|\mathbf{y}_{\text{obs}}) = \int \pi(\mathbf{y}_{\text{rep}}|\boldsymbol{\zeta})\ \pi(\boldsymbol{\zeta}|\mathbf{y}_{\text{obs}})\ d\boldsymbol{\zeta}, \tag{9}$$

is the predictive density of a new independent set of observables $\mathbf{y}_{\text{rep}}$ under the model given the actual data $\mathbf{y}_{\text{obs}}$. In what follows, we shall see that the posterior predictive distribution is easier to work with, because features of $\mathbf{y}_{\text{rep}}$ having the density (9) can easily be estimated when MCMC samples from the posterior $\pi(\boldsymbol{\zeta}|\mathbf{y}_{\text{obs}})$ are available.

The expected value of the loss function (8) with respect to the predictive distribution (9) is the proposed model selection criterion. This has attractive interpretations in terms of the classical likelihood ratio test statistic for comparing two models. Let the fitted probabilities (based on the mle) for a full and a reduced model be denoted by $\hat{\mathbf{P}}$ and $\tilde{\mathbf{P}}$, respectively. The likelihood ratio statistic for comparing the two models is given by

$$d(\hat{\mathbf{P}}, \tilde{\mathbf{P}}) = 2\sum_{r=1}^{s} y_{r,\text{obs}} \left(\log \hat{\mathbf{P}}_r - \log \tilde{\mathbf{P}}_r\right).$$

Note that if a so called saturated model is taken as the full model then $\hat{P}_r = y_{r,\text{obs}}/n$. Now the above statistic reduces to

$$d(\hat{\mathbf{P}}, \tilde{\mathbf{P}}) = 2\left(\sum_{r=1}^{s} y_{r,\text{obs}} \log \frac{y_{r,\text{obs}}}{n\,\tilde{\mathbf{P}}_r}\right). \tag{10}$$

This is the likelihood ratio test used by Bock and Aitkin (1981).

Let $P_r^* = E(y_{r,\text{rep}}/n)$ where the expectation is taken with respect to the predictive distribution (9). Now we have

$$\begin{aligned} E\left\{L(\mathbf{y}_{\text{rep}},\ \mathbf{y}_{\text{obs}})\right\} &= 2\textstyle\sum_{r=1}^{s} y_{r,\text{obs}}\Big[\log(y_{r,\text{obs}}/n) - E\left\{\log(y_{r,\text{rep}}/n)\right\}\Big] \\ &= 2\textstyle\sum_{r=1}^{s} y_{r,\text{obs}}\Big[\log(y_{r,\text{obs}}/n) - \log P_r^* + \log P_r^* - E\left\{\log(y_{r,\text{rep}}/n)\right\}\Big] \\ &= LRS + 2\textstyle\sum_{r=1}^{s} y_{r,\text{obs}}\left[\log P_r^* - E\left\{\log(y_{r,\text{rep}}/n)\right\}\right] \end{aligned}$$

where LRS is the likelihood ratio statistic (10) with the mle $\tilde{\mathbf{P}}$ replaced by $\mathbf{P}^*$. The LRS provides a goodness of fit measure as can be seen from its connection with the likelihood ratio test. Usually a more complex model provides a better fit and hence the LRS should go down when a more complex but nested model is fitted to the data set.

Using a Taylor series expansion for the log function the second term can be approximated as the variance of $y_{r,\text{rep}}$, see Waller *et al.* (1997) for a similar example. Hence the second term is likely to be high if the fitted model is too large for the data set. In other words, this takes care of uncertainty in estimation as the variability of $\mathbf{y}_{\text{rep}}$ is likely to be higher if a more complex model is fitted. Intuitively the parameters are less clearly identified hence more poorly estimated (i.e. variability is higher) under a more complex model. Henceforth, we use the following notation and decomposition for the expected loss function

$$EPD \equiv E\left\{L(\mathbf{y}_{\text{rep}},\ \mathbf{y}_{\text{obs}})\right\} = LRS + PEN, \tag{11}$$

where PEN is obtained by subtraction. The notation EPD stands for expected predictive deviance, another name for the expected loss function (8) used here.

Note the conflicting behavior of the two components in (11). As a more complex model is fitted the LRS should go down whereas the penalty should go up. Hence when fitting a sequence of more complex and nested models a trade-off must arise. At some intermediate model the increase in PEN will not be offset by the decrease in LRS. The model choice criterion (11) chooses this model as the best model for the data set. See the appendix for computational details for the quantities in (11).

4.3 A Simulation Example

We consider a simulation example to compare and study the sensitivities of the model choice criteria described above. We take $n = 200$ subjects with $k = 5$ items. The item difficulty parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$ are chosen as 5 equally spaced points in the interval [–3, 3]. The slope parameters are all set at 1. The subject parameter θ_i is simulated from its prior distribution which is the standard normal distribution. The guessing parameter c_j is also simulated from its prior distribution which in this case is the beta distribution with parameters 1 and 3. In each of the 1000 simulation replications we first generate a data set from the three parameter probit model.

We fit all three models to each simulated data set using the MCMC methods developed here. After an initial burn-in of 1000 iterations we use the next 5000 iterates to calculate all the model choice criteria discussed above. The whole experiment takes about a month to run on a 450 Mhz PC with 128 MB RAM running the linux operating system.

We report the results of the experiment in Table 2. The model choice criteria are given in rows and the three models are given in columns. Each entry in the table represents the number of times the particular model in the column is judged to be the best model according to the model choice criterion in the corresponding row.

Observe from the table that the EPD and the PsBF behave similarly; there is not much difference between the two rows. We have investigated also on the relative magnitude of each of the two criteria EPD and PsBF for the three different models. The absolute difference between the values of EPD (or PsBF) for two different models behaves like an exponential random variable. That is, for some data sets the EPD criterion (or PsBF) cannot discriminate well and for some other data sets the criterion strongly prefers one model against the other (corresponding to the right tail of the exponential distribution). Conflict between the two criteria arises in 374 out of 1000 replications. We have investigated the data sets corresponding to these 374 cases in detail. These did not reveal any obvious pattern for making general conclusions. However, in many of these cases we have observed that the absolute differences between the values of any of the two model choice criteria for two different models are not large. In other words the best model is not selected very strongly in these cases. In conclusion, this example finds that the two Bayesian predictive model selection criteria prefer the more complex two or three parameter models. In almost 50% of the simulations a three parameter model is selected as the best model.

	1PP	2PP	3PP
EPD	0	536	464
PsBF	2	523	475

Table 2: Model choice for the simulation example. An entry in the table indicates the number of replications for which the particular model (in the column) is judged to be the best according to the model choice criterion (corresponding to the row).

5 Law School Data Example

We consider a well known data set from Section 6 of the Law School Aptitude test (LSAT). Each of the 1000 candidates attempted 5 items in this section. The total number of correct responses for each item in order are 924, 709, 553, 763 and 870. There are 298 subjects who answered all 5 items correctly.

This data set has been extensively analyzed in the literature, see e.g. Baker (1992), Bock and Aitkin (1981). The `BUGS` software also fits one and two parameter models to this data set. We use the beta prior distribution with parameters 1 and 3 for the guessing parameter c_j. Empirical evidence based on the MCMC output suggests that we can work with a non-informative prior distribution for β_j. We take $\delta = 10^4$ and $\nu = 1$ which guarantees identifiability of the α_j.

We implement the Gibbs sampler for the logistic models using the ARMS as discussed in Section 3.1. For the normal-ogive models we undertake computation using the DAGS developed in Section 3.2. Several convergence diagnostics have been calculated and they did not show any particular sign of non-stationarity. We have used 10,000 iterates from the ARMS and 100,000 iterates from the DAGS to make inference.

Tables 3 and 4 provide the parameter estimates. The estimates of the item difficulty parameter (β_j) from all the models agree with the marginal totals of the correct responses. According to the estimates item 3 is the most difficult and item 1 is the easiest. The marginal totals agree with this. Item 3 is also the most discriminatory item and the estimates of α_3 point this out. The Bayes estimates and the Bock and Aitkin (1981) marginal maximum likelihood estimates for the logistic models are close.

	1PP	2PP		3PP		
Item	β	β	α	β	α	c
1	–0.82(0.06)	–0.70(0.08)	1.03(0.32)	–0.86(0.27)	0.85(0.32)	0.28(0.21)
2	0.30(0.04)	0.26(0.07)	1.07(0.29)	0.33(0.37)	1.22(0.51)	0.29(0.18)
3	0.84(0.04)	0.70(0.05)	1.38(0.60)	0.81(0.39)	1.51(0.66)	0.21(0.12)
4	0.10(0.04)	0.08(0.05)	0.97(0.25)	0.16(0.40)	1.11(0.50)	0.32(0.20)
5	–0.42(0.05)	–0.34(0.06)	0.86(0.28)	–0.43(0.33)	0.87(0.39)	0.30(0.21)

Table 3: Expected a posteriori estimates of the parameters of the normal-ogive models fitted to the LSAT data set. Standard deviations are given in parentheses.

Rasch		2PL				3PL		
β		β		α		β	α	c
MLE	Bayes	MLE	Bayes	MLE	Bayes	Bayes	Bayes	Bayes
–1.29	–1.32(0.10)	–1.30	–1.30(0.17)	1.10	1.07(0.34)	–1.52(0.52)	0.93(0.39)	0.28(0.20)
0.48	0.50(0.07)	0.48	0.48(0.09)	0.96	1.01(0.28)	0.56(0.62)	1.19(0.63)	0.27(0.18)
1.26	1.30(0.07)	1.22	1.22(0.08)	1.19	1.28(0.40)	1.44(0.66)	1.66(0.89)	0.21(0.12)
0.17	0.17(0.07)	0.19	0.19(0.10)	0.92	0.96(0.26)	0.13(0.56)	0.96(0.46)	0.26(0.19)
–0.63	–0.66(0.09)	–0.58	–0.59(0.12)	0.87	0.90(0.28)	–0.60(0.70)	0.92(0.47)	0.31(0.23)

Table 4: Parameter estimates for the parameters of the logistic models fitted to the LSAT data set. Standard deviations are given in parentheses. MLE stands for the Bock and Aitkin marginal ML estimates (MMLE) using the EM algorithm.

Table 5 shows the different model choice criteria for the two and three parameter models. The one parameter models are excluded because those did not provide adequate model fit. Also results from the simulation example in Section 4.3 justify this. The values of LRS are similar to those reported by Bock and Aitkin (1981) for the same data set.

According to the LRS the three parameter models provide better fit than the corresponding two parameter versions as expected. However, note that the penalty factor PEN is higher for the three parameter models. As a result the 2PL model is selected according to the EPD criterion. This is also confirmed by the PsBF criterion since the Bayes factor for the 2PL model is larger than 1 when compared against any other model. The CPOs for 2608 of the 5000 binary observations are higher for the 2PL model than the 3PL model. Hence the 2PL seems to be the best model for this data set, although it is not much better than the 2PP model.

	Normal-Ogive		Logistic	
	2PP	3PP	2PL	3PL
LRS	21.1	17.5	21.2	17.4
PEN	33.6	44.4	33.2	44.5
EPD	54.7	61.9	54.4	61.9
$ln(PSB)$	–2457.8	–2458.6	–2457.6	–2458.7

Table 5: Model choice criteria for different models fitted to the LSAT data set. LRS is the likelihood ratio statistic, PEN is the predictive variability penalty, EPD is the overall expected predictive deviance, and $ln(PSB)$ is needed to calculate the pseudo-Bayes Factor.

6 Discussion

This article proposes a new data augmentation scheme for running the Gibbs sampler for the three parameter item response models with probit link function. This extends the work of Albert (1992) to fit the three parameter models. Numerical examples show that the new scheme is much faster to run and is more efficient than the default Metropolis scheme.

Another contribution of this article is the investigation into model selection procedures. A simulation investigation shows that the two predictive Bayesian model selection criteria prefer more complex two and three parameter models. It also gives justification for considering the more complex three parameter models for which a new MCMC computing method has been developed in this article.

Appendix: Details for computing the model choice criteria

Computing the PsBF

To compute the PsBF we have to evaluate the CPO for each binary response. Note that here $\mathbf{y}_{\mathrm{obs}}$ is the vector of nk $(= N)$ binary observations and $\boldsymbol{\zeta}$ is the vector of all parameters under the assumed model. Further, it is obvious that under any of the models the responses $y_{ij}, i = 1, \ldots, n; j = 1, \ldots, k$ are conditionally independent given all the parameters. In this situation it is relatively straightforward to evaluate the CPO for the ijth observation y_{ij}. Suppose that $\boldsymbol{\zeta}^{(1)}, \ldots, \boldsymbol{\zeta}^{(B)}$ denote B Gibbs sampled values from $\pi(\boldsymbol{\zeta}|\mathbf{y}_{\mathrm{obs}})$. A Monte Carlo estimate of $\pi(y_{ij}|\mathbf{y}_{(\mathrm{r}),\mathrm{obs}})$ is

$$\hat{\pi}(y_{ij}|\mathbf{y}_{(\mathrm{r}),\mathrm{obs}}) = \left(\frac{1}{B} \sum_{t=1}^{B} \frac{1}{p_{ij}^{y_{ij}} (1 - p_{ij})^{1-y_{ij}}} \right)^{-1},$$

where p_{ij} is the probability of the correct response under the assumed model and it is evaluated at the simulated parameter values $\boldsymbol{\zeta}^{(t)}$. In other words, the CPO is estimated by the harmonic mean of the Bernoulli probability mass functions.

Computing the EPD criterion

We now provide the computational details to approximate criterion (11). We first give details to estimate the unconditional probabilities P_r for the rth pattern under the assumed model. Without loss of generality assume that the ith person $(i = 1, \ldots, n)$ with given ability θ has the rth score pattern. Let $P_{r|\theta}$ be the conditional probability of this event. If we have fitted a model for the probabilities p_{ij} then

$$P_{r|\theta} = \prod_{j=1}^{k} p_{ij}^{y_{ij}} (1 - p_{ij})^{1-y_{ij}} \tag{12}$$

where θ_i (in the definition of p_{ij}) is replaced by θ. Let $g(\theta)$ denote the prior ability distribution. Now the unconditional multinomial probabilities $P_r, r = 1, \ldots, s$ is given by $P_r = \int P_{r|\theta}\; g(\theta)\; d\theta$, see e.g. Bock and Aitkin (1981). This integral can be easily estimated using Monte Carlo integration as follows. At the tth iteration we sample a new ability value $\theta^{(t)}$ from the standard normal distribution. Then we calculate $P_{r|\theta^{(t)}}$ using the model and equation (12). Now the sample average

$$\frac{1}{B}\sum_{t=1}^{B} P_{r|\theta^{(t)}} \tag{13}$$

is an estimate of P_r.

Calculation of the expected loss function (11) however, is more involved. This requires estimating the probabilities P_r at each MCMC iteration rather than at the end of the MCMC run. We overcome this problem by generating a number, C say, of new ability values $\theta^{(1)}, \dots, \theta^{(C)}$ at each iteration t. Then we form the average $\frac{1}{C}\sum_{i=1}^{C} P_{r|\theta^{(i)}}$ to estimate P_r. Empirical evidence for the example in Section 5 suggests that a value as small as 10 is good enough because the estimates do not change very much if we take a larger value.

Note that under the posterior predictive density (9) the replicated data $\mathbf{y}_{\text{rep}}$ is a multinomial observation with parameters n and the estimated probability vector $\mathbf{P}$. At each iteration we obtain a new multinomial observation $\mathbf{y}_{\text{rep}}$ and evaluate the loss function (8). Average of these evaluations at the end of the MCMC run is an estimate of EPD. Also after the MCMC run we evaluate the probability P_r using (13). These probabilities are then put back in (10) to obtain LRS. PEN is obtained by subtraction.

REFERENCES

Aitkin, M. (1997) The calibration of P-values, posterior Bayes factors and the AIC from the posterior distribution of the likelihood. *Statistics and Computing,* **7**, 253–261.

Albert, J. H. (1992) Bayesian Estimation of Normal Ogive Item Response Curves Using Gibbs Sampling. *J. Educational Statist.*, **17**, 251–269.

Albert, J. H. and Chib, S. (1993) Bayesian Analysis of Binary and Polytomous Response Data. *J. Amer. Staist. Assoc.*, **88**, 669–679.

Andersen, E. B. (1973) A goodness of fit test for the Rasch model. *Psychometrika*, **38**, 123–140.

Baker, F. B. (1992) *Item Response Theory.* New York: Marcel Dekker.

Birnbaum, A. (1968) Some latent trait models and their use in inferring an examinee's ability. In *Statistical Theories of Mental Test Scores.* (Eds. F. M. Lord and M. R. Novick). Reading: Addison-Wesley, pp 397–472.

Bock, R. D. and Aitkin, M. (1981) Marginal Maximum Likelihood Estimation of Item Parameters: Application of an EM Algorithm. *Psychometrika*, **46**, 443–459.

Geisser, S. and Eddy, W. (1979) A predictive approach to model selection. *J. Amer. Staist. Assoc.*, **74**, 153–160.

Gelfand, A. E. (1996) Model determination using sampling based methods. In *Markov Chain Monte Carlo in Practice.* (Eds. W. R. Gilks, S. Richardson and D. J. Spiegelhalter). London: Chapman and Hall, pp 145–161.

Gelfand, A. E. and Smith, A. F. M. (1990) Sampling based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.*, **85**, 398–409.

Ghosh, M., Ghosh., A., Chen, M.H. and Agresti, A. (2000) Noninformative priors for one-parameter item response models. *J. Stat. Plan. Inf.*, **88**, 99–115.

Gilks, W. R., Best, N. G. and Tan, K. K. C. (1995) Adaptive Rejection Metropoils Sampling within Gibbs Sampling. *Appl. Statist.*, 455–472.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. G. (1996) *Markov Chain Monte Carlo In Practice.* London: Chapman and Hall.

Gilks, W.R. and Wild, P. (1992) Adaptive rejection sampling for Gibbs sampling. *Appl. Statist.*, **41**, 337–348.

Johnson, V. E. and Albert, J. H. (1999) *Ordinal Data Modeling.* New York: Springer.

Kass, R. E., Carlin, B. P., Gelman, A. and Neal, R. (1998) Markov chain Monte Carlo in practice: A roundtable discussion. *Amer. Stat.*, **52**, 93–100.

Kass, R. E. and Raftery, A. E. (1995) Bayes factors. *J. Amer. Statist. Assoc.*, **90**, 773–795.

Key, J. T., Pericchi, L. R. and Smith, A. F. M. (1999) Bayesian Model Choice: What and Why? In *Bayesian Statistics 6* (Eds. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith). Oxford University Press, to appear.

Lord, F. M. (1975) Relative efficiency of number-right and formula scores. *British Journal of Mathematical and Statistical Psycology*, **28**, 46–50.

Lord, F. M. (1980) *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: Erlbaum.

Mislevy, R. J. amd Bock, R. D. (1986) *PC-BILOG: Item analysis and test scoring with binary logistic models.* Mooresville, IN: Scientific Software Inc.

Novick, M. R., Jackson, P. H., Thayer, D. T. and Cole, N. S. (1972) Estimating multiple regression in m groups: A cross-validation study. *British Journal of Mathematical and Statistical Psycology*, **5**, 33–50.

Patz, R. J. and Junker, B. W. (1999a) A straightforward approach to Markov chain Monte Carlo methods for item response models. *J. Educ. Behav. Stat.*, **24**, 146–178.

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

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

Rasch, G. (1961) On general laws and the meaning of the measurement in psychology, Vol 4, (pp. 321–334). In *Proceedings of the 4th Berkley Symposium on Mathematical Statistics.* Berkley: University of California Press.

Roberts, G. O. (1996) Markov chain concepts related to sampling algorithms. In *Markov Chain Monte Carlo in Practice.* (Eds. W. R. Gilks, S. Richardson and D. J. Spiegelhalter). London: Chapman and Hall, pp 45–57.

Sahu, S. K. and Roberts, G. O. (1999) On convergence of the EM algorithm and the Gibbs sampler. *Statistics and Computing*, **9**, 55–64.

Spiegelhalter, D. J., Best, N. G. and Carlin, B. P. (1998) Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. *Technical Report*, MRC Biostatistics Unit, Cambridge.

Spiegelhalter, D. J., Thomas, A. and Best, N. G. (1996) Computation on Bayesian graphical models. In *Bayesian Statistics 5*, (Eds. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith). Oxford: Oxford University Press, pp. 407–426.

Swaminathan, H. and Gifford, J. A. (1986) Bayesian Estimation in the Three Parameter Logistic Model. *Psychometrika*, **51**, 589–601.

Tsutakawa, R. K. and Lin, H. Y. (1986) Bayesian Estimation of Item Response Curves. *Psychometrika*, **51**, 251–267.

Tanner, M. A. (1996) *Tools for Statistical Inference.* Springer-Verlag: Heidelberg.

van der Linden, W. and Hambleton, R. K. (1997) *Handbook of Modern Item Response Theory*, (eds). New York: Springer.

Waller, L. A., Carlin, B. P., Xia, H. and Gelfand, A. E. (1997) Hierarchical Spatio-Temporal Mapping of Disease Rates. *J. Amer. Statist. Assoc.*, **92**, 607–617.