PSYCHOMETRIKA—VOL. 65, NO. 4, 475–496
DECEMBER 2000

BAYESIAN FACTOR ANALYSIS FOR MULTILEVEL BINARY OBSERVATIONS

ASIM ANSARI AND KAMEL JEDIDI

COLUMBIA UNIVERSITY

Multilevel covariance structure models have become increasingly popular in the psychometric literature in the past few years to account for population heterogeneity and complex study designs. We develop practical simulation based procedures for Bayesian inference of multilevel binary factor analysis models. We illustrate how Markov Chain Monte Carlo procedures such as Gibbs sampling and Metropolis–Hastings methods can be used to perform Bayesian inference, model checking and model comparison without the need for multidimensional numerical integration. We illustrate the proposed estimation methods using three simulation studies and an application involving student's achievement results in different areas of mathematics.

Key words: MCMC methods, binary data, multilevel factor analysis, Gibbs sampling, Metropolis–Hastings.

1. Introduction

Multilevel covariance structure models have become increasingly popular in the psychometric literature in the past few years (Goldstein & McDonald, 1988; Longford & Muthén, 1992; McDonald & Goldstein, 1989; Muthén, 1989, 1994; Muthén & Satorra, 1989). The rapid growth of multilevel modeling reflects the realization that it is crucial to account for population heterogeneity in order to make valid inferences from data that have a nested or hierarchical structure. Such nested data structures are common in studies of student achievements in classrooms and schools. Here students can be considered nested within classrooms and classrooms can be considered nested within schools thus forming a three-level nesting structure. The existing literature on multilevel latent variable models mostly deals with continuous observed variables. In this paper we develop hierarchical Bayesian methods for performing factor analysis of multilevel binary data. We extend previous work on multilevel factor analysis (Longford & Muthén) by focusing on data containing dichotomous observed variables. Although we concentrate on data involving only binary variables, our procedures can also handle metric data and mixed (i.e., metric and binary) data situations with minor modifications.

There is a rich tradition on covariance structure modeling of binary data. Primary contributions in this area have come from Christofferson (1975), Bartholomew (1980, 1981, 1984), Muthén and Christofferson (1981), Muthén (1978, 1984, 1987), Bock and Aitken (1981) and Bock and Gibbons (1996). Muthén (1984, 1987) presented a general computer program LISCOMP for performing latent variable modeling of binary data. Bock and Gibbons developed a maximum likelihood approach for exploratory factor analysis and used a combination of the EM algorithm and scoring methods to estimate the model parameters. Maximum likelihood approaches to factor analysis of binary data pose difficulty in the estimation of model parameters as they requires multidimensional numerical integration. This difficulty is further compounded with *multilevel* data structures. In this paper we describe the use of Markov Chain Monte Carlo (MCMC) procedures for simulation based estimation of factor analysis models. These procedures circumvent the need for evaluating multidimensional integrals and are therefore eminently suitable for binary multilevel data situations. We also discuss how considerations of model com-

The authors thank Ian Westbury, University of Illinois at Urbana Champaign for kindly providing the SIMS data for the application.

Requests for reprints should be sent to Asim Ansari, 517 Uris Hall, Columbia University, 3022 Broadway, New York, NY, 10027. E-mail: maa48@columbia.edu

0033-3123/2000-4/1998-0639-A $00.75/0

parison and model adequacy of binary factor analysis models can be handled using the simulation output from MCMC procedures.

Previous research in the psychometric literature has used Bayesian methods in factor analysis models for two distinct purposes. The early work of Martin and McDonald (1975) illustrated the use of Bayesian techniques for factor analysis of continuous variables to circumvent the problem of Heywood cases. Lee (1981) illustrated the use of Bayesian confirmatory factor analysis under different forms of prior distributions. The second stream of research has focused on procedures illustrating how point estimates of model parameters can be used for making Bayesian posterior analysis of factor scores (see Bartholomew, 1981; and Shi & Lee, 1997). These procedures, however, ignore the uncertainty pertaining to the other model parameters. In contrast, our analysis procedures permit the direct estimation of factor scores along with the other model parameters and therefore allow a proper accounting of uncertainty in making inferences regarding all unknown quantities in the model. In addition, Bayesian procedures do not rely on asymptotic inference and can be especially useful in nonlinear models (Arminger & Muthén, 1998) and binary data situations as these may require very large sample sizes for asymptotic properties to hold.

The rest of the paper is organized as follows. Section 2 presents a two-level factor analysis model and discusses identification and the specification of priors. Section 3 outlines the MCMC algorithm for estimation and provides a description of the full conditional distributions. Section 4 discusses procedures for model comparison and adequacy. Section 5 presents the results of three simulation studies. Section 6 illustrates the model using mathematics achievement data and Section 7 concludes with a discussion of limitations and opportunities for future research.

2. Model

Multilevel covariance structure modeling assumes that data are obtained by cluster sampling, that is, by randomly sampling the units at each level of the hierarchy. For example, researchers in education first randomly sample a subset of classrooms and then select a random sample of students within each selected class to obtain a two-level data structure. This sampling scheme therefore requires a two-level model specification. The first level captures the within group (i.e, student variation within a classroom) while the second level models the between group (i.e, across classes) variation.

In this section we describe a two-level binary factor analysis model. Suppose data come from I distinct groups (e.g., classrooms) indexed $i = 1$ to I. Each group i, provides $j = 1$ to n_i observations (e.g., student responses) on a p dimensional vector $\mathbf{y}_{ij}$ of binary random variables. The total number of observations in the two-level data is then given by $N = \sum_i n_i$. The observed binary variables can be modeled in terms of p underlying continuous variables $\mathbf{w}_{ij}$ that have an interpretation which depends on the context of the application. In psychometric studies dealing with achievement data, the latent variables refer to underlying ability variables. In biometric applications these describe tolerances while in consumer psychology studies they refer to unobserved utility for products. The link between the observed binary variables and the underlying latent variables for observation j of group i can be represented in terms of a threshold specification as follows:

$$y_{ijk} = \begin{cases} 1 & \text{if } w_{ijk} > 0 \\ 0 & \text{otherwise,} \end{cases} \qquad \text{for } k = 1 \text{ to } p. \tag{1}$$

At the first level, we assume that conditional on the group mean $\boldsymbol{\tau}_i$, the latent variables within a group i have a common factor structure:

$$\begin{aligned} \mathbf{w}_{ij} &\sim N_p(\boldsymbol{\tau}_i, \mathbf{V}_1), \quad \text{(iid)} \\ \mathbf{V}_1 &= \boldsymbol{\Omega}_1 \boldsymbol{\Psi}_1 \boldsymbol{\Omega}_1' + \boldsymbol{\Delta}_1 \end{aligned} \tag{2}$$

where $\mathbf{\Omega}_1$ is a $p \times r_1$ matrix of factor loadings ($r_1 \leq p$), $\mathbf{\Psi}_1$ is a $r_1 \times r_1$ positive definite variance matrix and $\mathbf{\Delta}_1$ is a diagonal matrix of variances. At the second level, the group means τ_i can be assumed to be normally distributed, that is,

$$\begin{aligned} \boldsymbol{\tau}_i &\sim N_p(\boldsymbol{\nu}, \mathbf{V}_2), \quad \text{(iid)} \\ \mathbf{V}_2 &= \mathbf{\Omega}_2 \mathbf{\Psi}_2 \mathbf{\Omega}_2' + \mathbf{\Delta}_2. \end{aligned} \tag{3}$$

Here $\mathbf{\Omega}_2$ is a loading matrix of dimension $p \times r_2$, $\mathbf{\Psi}_2$ is a covariance matrix of dimensions $r_2 \times r_2$ and $\mathbf{\Delta}_2$ is a diagonal covariance matrix. The two-level model can alternatively be described in terms of factor scores as follows:

$$\begin{aligned} \mathbf{w}_{ij} &= \boldsymbol{\tau}_i + \mathbf{\Omega}_1 \boldsymbol{\delta}_{1,ij} + \boldsymbol{\epsilon}_{1,ij} \\ \boldsymbol{\tau}_i &= \boldsymbol{\nu} + \mathbf{\Omega}_2 \boldsymbol{\delta}_{2,i} + \boldsymbol{\epsilon}_{2,i} \end{aligned} \tag{4}$$

where $\boldsymbol{\delta}_{1,ij} \sim N(\mathbf{0}_{r_1}, \mathbf{\Psi}_1)$, $\boldsymbol{\delta}_{2,i} \sim N(\mathbf{0}_{r_2}, \mathbf{\Psi}_2)$, $\boldsymbol{\epsilon}_{1,ij} \sim N(\mathbf{0}_p, \mathbf{\Delta}_1)$, and $\boldsymbol{\epsilon}_{2,i} \sim N(\mathbf{0}_p, \mathbf{\Delta}_2)$ are mutually independent normal random vectors. The vectors $\boldsymbol{\delta}_{1,ij}$ and $\boldsymbol{\delta}_{2,i}$ represent the first level and second level factor scores, respectively. Notice that if $\mathbf{w}_{ij}$ are observed then we obtain the special case of a multilevel continuous factor analysis model.

2.1. *Identification*

We begin discussing identification of the two-level model by focusing on the scale restrictions imposed by the binary nature of the observed variables. Muthén (1979) discusses identification in general covariance structure models for binary data while Chib and Greenberg (1998) discuss identification for multivariate probit regression models. The model for group i is given by $\mathbf{w}_{ij} = \tau_i + \mathbf{e}_{1,ij}$, where $\mathbf{e}_{1,ij} \sim N(0, \mathbf{V}_1)$. The binary nature of the observed variables implies that $\mathbf{V}_1$ cannot be estimated as an unrestricted covariance matrix. It is clear that the observed binary outcomes do not change if each of the p latent variables in $\mathbf{w}_{ij}$ is multiplied by a positive constant. Collecting these constants in a diagonal transformation matrix $\mathbf{T} = \text{diag}(t_1, \ldots, t_p)$, we see that the above system of equations is indistinguishable from the system $\mathbf{T}\mathbf{w}_i = \mathbf{T}\boldsymbol{\tau}_i + \mathbf{T}\mathbf{e}_{1,ij}$. In other words, the data cannot distinguish between the variances $\mathbf{V}_1$ and $\mathbf{T}\mathbf{V}_1\mathbf{T}$. However, as is well known, this problem can be fixed by choosing $\mathbf{T} = \text{diag}(v_{1,11}^{-1/2}, \ldots, v_{1,pp}^{-1/2})$, where $v_{1,kk}$ is the k-th diagonal element of $\mathbf{V}_1$. Then $\mathbf{T}\mathbf{V}_1\mathbf{T}$ reduces to a correlation matrix $\mathbf{\Sigma}_1$. This scaling transforms the vector $\mathbf{w}_{ij}$ to $\mathbf{u}_{ij} = \mathbf{T}\mathbf{w}_{ij}$ and the group means $\boldsymbol{\tau}_i$ to $\mathbf{m}_i = \mathbf{T}\boldsymbol{\tau}_i$. We therefore have a scaled level-one model given by

$$\begin{aligned} \mathbf{u}_{ij} &\sim N_p(\mathbf{m}_i, \mathbf{\Sigma}_1), \quad \text{(iid)} \\ \mathbf{\Sigma}_1 &= \mathbf{\Lambda}_1 \mathbf{\Psi}_1 \mathbf{\Lambda}_1' + \mathbf{\Theta}_1 \end{aligned} \tag{5}$$

where $\mathbf{\Lambda}_1 = \mathbf{T}\mathbf{\Omega}_1$ and $\mathbf{\Theta}_1 = \mathbf{T}\mathbf{\Delta}_1\mathbf{T}$. As the diagonal elements of $\mathbf{\Sigma}_1$ are fixed to unity, the elements of $\mathbf{\Theta}_1$ are given by the identity

$$\mathbf{\Theta}_1 = \mathbf{I} - \text{diag}(\mathbf{\Lambda}_1 \mathbf{\Psi}_1 \mathbf{\Lambda}_1'). \tag{6}$$

At the second level of the hierarchy, the scaled group means $\mathbf{m}_i$ can be now represented as

$$\begin{aligned} \mathbf{m}_i &\sim N_p(\boldsymbol{\mu}, \mathbf{\Sigma}_2), \quad \text{(iid)} \\ \mathbf{\Sigma}_2 &= \mathbf{\Lambda}_2 \mathbf{\Psi}_2 \mathbf{\Lambda}_2' + \mathbf{\Theta}_2, \end{aligned} \tag{7}$$

where $\mathbf{\Lambda}_2 = \mathbf{T}\mathbf{\Omega}_2$, $\boldsymbol{\mu} = \mathbf{T}\boldsymbol{\nu}$ and $\mathbf{\Theta}_2 = \mathbf{T}\mathbf{\Delta}_2\mathbf{T}$. It is important to note that the rescaled variance matrix $\mathbf{\Sigma}_2$ is not a correlation matrix. The diagonal elements of $\mathbf{\Sigma}_2$ give the ratio of the between

group and within group variances. Further analysis in this paper will be based on the identified set of parameters.

In addition to the scaling restrictions discussed above, further restrictions are required that depend upon whether we have a confirmatory or an exploratory model. In confirmatory models, the loadings matrices, $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$, have certain elements restricted to zero. Furthermore, in order to fix the scale of the latent factors, one can either impose restrictions via the loadings matrices (e.g., set the scale of the factor to the scale of an *a priori* chosen variable), or can assume that $\mathbf{\Psi}_1$ and $\mathbf{\Psi}_2$ are correlation matrices. In this paper, we adopt the latter restriction. In exploratory factor analysis, $\mathbf{\Psi}_1$ and $\mathbf{\Psi}_2$ are typically assumed to be identity matrices. In addition, we need to impose $\frac{r_1(r_1-1)}{2}$ restrictions on $\mathbf{\Lambda}_1$ and $\frac{r_2(r_2-1)}{2}$ restrictions on $\mathbf{\Lambda}_2$ to account for rotational indeterminacies. Bock and Gibbons (1996) suggest one form in which these restrictions can be applied. In the rest of the paper we will focus on the confirmatory factor analysis model. Extensions to the exploratory case are straightforward[1].

2.2. *Priors*

Bayesian procedures require the specification of priors for $\boldsymbol{\gamma} = \{\boldsymbol{\mu}, \mathbf{\Lambda}_2, \mathbf{\Theta}_2, \mathbf{\Psi}_2, \mathbf{\Lambda}_1, \mathbf{\Psi}_1\}$. Lee (1981) discusses different forms of prior distributions for continuous factor analysis models. In this paper we specify the prior distribution as follows:

$$p(\boldsymbol{\gamma}) = p(\boldsymbol{\mu})p(\mathbf{\Theta}_2)p(\mathbf{\Lambda}_2)p(\mathbf{\Psi}_2)p(\mathbf{\Lambda}_1)p(\mathbf{\Psi}_1)I(S(\mathbf{\Lambda}_1, \mathbf{\Psi}_1)). \tag{8}$$

We use proper but diffuse priors over model parameters. The prior for the overall mean $\boldsymbol{\mu}$ can be chosen as a multivariate normal distribution $N(\boldsymbol{\eta}, \mathbf{C})$. The covariance matrix $\mathbf{C}$ in this prior can be assumed to be diagonal. The diagonal elements (variances) can be set to large values to represent vague knowledge. The exact location $\boldsymbol{\eta}$ is no longer critical when $\mathbf{C}$ is large and $\boldsymbol{\eta}$ can, therefore, be set to a zero.

The matrix $\mathbf{\Theta}_2$ is a $p \times p$ diagonal matrix containing the measurement error variances. In keeping with standard Bayesian analysis, we can assume independent inverse gamma priors over the variances. We therefore have $IG(a_k, b_k)$ for the k-th variance $\theta_{2,kk}$. The constants a_k and b_k can be chosen to ensure a vague but proper prior.

The correlation matrix $\mathbf{\Psi}_2$ has $r_{f2} = r_2(r_2 - 1)/2$ nonredundant correlations which are the only free parameters of the matrix. Let vec($\mathbf{\Psi}_2$) be a vector of these free correlations. Following Chib and Greenberg (1998), we assume a multivariate normal prior over vec($\mathbf{\Psi}_2$). Formally we have

$$p(\text{vec}(\mathbf{\Psi}_2)) \propto \exp\left[-\frac{1}{2}(\text{vec}(\mathbf{\Psi}_2) - \text{vec}(\mathbf{\Psi}_{2,0}))' \mathbf{G}_{2,0}(\text{vec}(\mathbf{\Psi}_2) - \text{vec}(\mathbf{\Psi}_{2,0}))\right], \tag{9}$$

where vec($\mathbf{\Psi}_2$) belongs to a subset of the hypercube $[-1, 1]^{r_{f2}}$ that leads to a proper correlation matrix. If we do not have strong prior information about the likely magnitude of the correlations, we can choose vec($\mathbf{\Psi}_{2,0}$) to be a null vector of dimension r_{f2} and the precision matrix $\mathbf{G}_{2,0}$ can be conveniently specified as a $r_{f2} \times r_{f2}$ identity matrix.

The matrix $\mathbf{\Lambda}_2$ has a patterned structure. We therefore specify independent multivariate normal priors over the nonzero elements within each row of the matrix. We have for row k a prior $N(\mathbf{g}_{2k}, \mathbf{H}_{2k})$. The covariance matrix $\mathbf{H}_{2k}$ can be assumed diagonal with large variances to ensure a diffuse prior. The prior over $\mathbf{\Lambda}_2$ then is a product of the independent priors associated with the p rows.

The prior for the loadings in the level-one matrix $\mathbf{\Lambda}_1$ can be specified in an analogous manner. We need to specify priors only on the unrestricted elements of this matrix. We therefore

[1] As suggested by an anonymous reviewer, it is possible to impose cross-level constraints such as $\mathbf{\Lambda}_1 = \mathbf{\Lambda}_2$ when the factor structure is *common* across levels. Here we consider the more general situation that allows the number of factors to be different across levels.

use independent multivariate normal priors $N(\mathbf{g}_{1k}, \mathbf{H}_{1k})$ over the unrestricted elements of each vector $\boldsymbol{\lambda}_{1k}$, $k = 1$ to p, where $\boldsymbol{\lambda}_{1k}'$ is row k of $\boldsymbol{\Lambda}_1$. The prior over $\boldsymbol{\Lambda}_1$ then is a product of the independent priors associated with the p rows.

The correlation matrix, $\boldsymbol{\Psi}_1$ has $r_{f1} = r_1(r_1 - 1)/2$ nonredundant correlations. Let vec($\boldsymbol{\Psi}_1$) be a vector of these free correlations. Analogous to the prior for $\boldsymbol{\Psi}_2$, we specify a multivariate normal prior over vec($\boldsymbol{\Psi}_1$). We have

$$p(\text{vec}(\boldsymbol{\Psi}_1)) \propto \exp\left[-\frac{1}{2}(\text{vec}(\boldsymbol{\Psi}_1) - \text{vec}(\boldsymbol{\Psi}_{1,0}))' \mathbf{G}_{1,0}(\text{vec}(\boldsymbol{\Psi}_1) - \text{vec}(\boldsymbol{\Psi}_{1,0}))\right], \tag{10}$$

where vec($\boldsymbol{\Psi}_1$) belongs to a subset of the hypercube $[-1, 1]^{r_{f1}}$ that leads to a proper correlation matrix. Once again we can choose vec($\boldsymbol{\Psi}_{1,0}$) to be a null vector of dimension r_{f1} and the precision matrix $\mathbf{G}_{1,0}$ can be specified as a $r_{f1} \times r_{f1}$ identity matrix.

The last term in the joint prior, $I(S(\boldsymbol{\Lambda}_1, \boldsymbol{\Psi}_1))$, accounts for the identification restrictions imposed by the fact that $\boldsymbol{\Sigma}_1$ is a correlation matrix. As $\boldsymbol{\Sigma}_1 = \boldsymbol{\Lambda}_1\boldsymbol{\Psi}_1\boldsymbol{\Lambda}_1' + \boldsymbol{\Theta}_1$, we need to ensure that $\mathbf{I} - \text{diag}(\boldsymbol{\Lambda}_1\boldsymbol{\Psi}_1\boldsymbol{\Lambda}_1') > 0$ for $\boldsymbol{\Theta}_1$ to be positive definite. Let S be the set of parameters $\{\boldsymbol{\Lambda}_1, \boldsymbol{\Psi}_1\}$ that satisfy the above inequality. Then $I(S(\boldsymbol{\Lambda}_1, \boldsymbol{\Psi}_1))$ is an indicator function for the set $S(\boldsymbol{\Lambda}_1, \boldsymbol{\Psi}_1)$ and is required to restrict the support of the joint prior density to values that satisfy the inequality.

3. Inference Procedure

The conditional probability of observing a binary vector $\mathbf{y}_{ij}$ for observation j of group i given the group mean $\mathbf{m}_i$ can be written as

$$g(\mathbf{y}_{ij} \mid \mathbf{m}_i) = \int_{s_1}\int_{s_2}\cdots\int_{s_p} f(\mathbf{u})d\mathbf{u}, \tag{11}$$

where $f(\mathbf{u})$ is a multivariate normal density specified in (5). The limits of integration are defined as follows:

$$S_k = \begin{cases} (-\infty, 0), & \text{if } y_{ijk} = 0 \\ (0, \infty), & \text{if } y_{ijk} = 1. \end{cases} \tag{12}$$

Given a sample of n_i independent observations $\mathbf{y}_{i1}, \mathbf{y}_{i2}, \ldots, \mathbf{y}_{in_i}$, the conditional likelihood for group i is given by the product of the n_i multiple integrals of the type given in (11), i.e.,

$$L_i(\boldsymbol{\Lambda}_1, \boldsymbol{\Psi}_1 \mid \mathbf{m}_i, \{\mathbf{y}_{ij}\}) = \prod_{j=1}^{n_i} g(\mathbf{y}_{ij} \mid \mathbf{m}_i). \tag{13}$$

The likelihood for the entire sample can be constructed as the product of the unconditional likelihoods for the I groups, that is,

$$L(\boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2, \mu, \boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2, \boldsymbol{\Theta}_2; \{\{\mathbf{y}_{ij}\}\}) = \prod_{i=1}^{I}\int\int\cdots\int L_i(\boldsymbol{\Lambda}_1, \boldsymbol{\Psi}_1 \mid \mathbf{m}_i, \{\mathbf{y}_{ij}\})h(\mathbf{m}_i)d\mathbf{m}_i, \tag{14}$$

where $h(\mathbf{m})$ is a multivariate normal density given in (7).

The above likelihood expression involves computation of high order multidimensional integrals and this makes classical inference based on maximum likelihood methods extremely difficult. In the Bayesian framework, inference about the unknown parameters is based on their joint posterior distribution. As the posterior density of all unknowns is very complex we use simulation based methods for summarizing the posterior density. This involves generating many random draws of the unknowns from the joint posterior density and then basing inference on the empirical distribution of this sample of draws. In our model, the complexity of the posterior

density precludes the use of direct methods for obtaining parameter draws from the posterior. We therefore use Markov Chain Monte Carlo (MCMC) methods to simulate the draws. Specifically, we use substitution sampling (a combination of the Gibbs sampler and the Metropolis Hastings algorithm) in tandem with data augmentation (Albert & Chib, 1993; Tanner & Wong, 1987) to obtain the requisite sample of parameter draws from the joint posterior distribution.

Substitution sampling involves replacing one complicated draw from the joint posterior with a sequence of relatively simple draws from easy to sample distributions. Sampling from the posterior is usually achieved by sampling from the full conditional distributions of blocks of parameters. If all the full conditional distributions are known, then substitution sampling reduces to a procedure known as Gibbs Sampling (Gelfand & Smith 1990; Geman & Geman, 1984). When the full conditional distribution for some parameter is not completely known (i.e., known only upto a normalizing constant) the Gibbs sampling step cannot be used and is therefore replaced by a Metropolis–Hastings step (Chib & Greenberg, 1995). In the context of the model developed in this paper we need to generate random draws for the unknowns, $\{\mathbf{\Lambda}_1, \mathbf{\Lambda}_2, \boldsymbol{\mu}, \{\mathbf{m}_i\}, \mathbf{\Psi}_1, \mathbf{\Psi}_2, \mathbf{\Theta}_2, \{\boldsymbol{\delta}_{1,ij}\}, \{\boldsymbol{\delta}_{2,i}\}, \{\{\mathbf{u}_{ij}\}\}\}$. Each iteration of the substitution sampler involves sequentially sampling from the full conditional distributions associated with each block of parameters. It is important to point out that the MCMC procedures described in this paper produces samples for the unknown $\{\{\mathbf{u}_{ij}\}\}$, via data augmentation (Albert & Chib, 1993; Tanner & Wong, 1987) and therefore circumvents the need for integration procedures. In addition, the MCMC procedures provide samples of the factor scores $\{\{\boldsymbol{\delta}_{1,ij}\}\}$, and $\{\boldsymbol{\delta}_{2,i}\}$, thus enabling posterior inference about the factor scores. Treating the factor scores as part of the unknowns in the model facilitates proper accounting of uncertainty regarding these quantities.

The substitution sampler is run for a large number of iterations. This iterative scheme of sequential draws generates a Markov chain that converges in distribution to the joint posterior under fairly general conditions (Tierney, 1994). After passing through an initial transient phase the chain converges to the posterior distribution of parameters and therefore the subsequent draws from the chain can be regarded as a sample from the posterior distribution. Geyer (1992) recommends a single long run to obtain a sample from the posterior, whereas Gelman and Rubin (1992) propose multiple chains from different starting values to help diagnose convergence. While convergence cannot be proved, a number of convergence diagnostics which use the statistical properties of the chain have been proposed in the literature. Cowles and Carlin (1996) and Brooks and Roberts (in press) provide detailed reviews of many of the methods proposed in the literature. After the chains have converged, a large sample of draws can be obtained to approximate the posterior distribution to any desired degree of accuracy. The Appendix describes the full conditional distributions and the simulation steps involved in each iteration of the Markov chain.

4. Model Assessment

4.1. Model Adequacy

The adequacy of a Bayesian model can be assessed using posterior predictive model checking (Gelman et al., 1996). Let y^{obs} be the observed data and γ be the vector of all unknowns. The sample of parameter draws $\gamma_1, \gamma_2, \ldots \gamma_d$ available from the MCMC algorithm can be used along with the appropriate sampling distribution $p(y \mid \gamma)$ to generate hypothetical replicated multilevel data sets $y_1^{rep}, y_2^{rep}, \ldots, y_d^{rep}$. The actual data set can be compared with the replicated data sets using test quantities $T(y, \gamma)$ involving either the data alone or both data and parameters. These test quantities are chosen to measure departures of the observed data from the assumed model. They can be omnibus goodness of fit measures or could be chosen specifically to highlight substantive aspects of the application of interest. If the replicated data sets differ systematically from the actual data on some test quantities, then we can ascertain that the model does not adequately capture the data generation process on those aspects that are captured by the test quantities.

A posterior predictive p value given by

$$p(y) = P[T(y^{rep}, \gamma) \geq T(y^{obs}, \gamma) \mid y^{obs}] \tag{15}$$

can be used to detect model inadequacies. This p-value can be approximated easily from the MCMC sequence of draws using

$$p(y) = \frac{1}{d} \sum_{i=1}^{d} I(T(y_i^{rep}, \gamma_i) \geq T(y^{obs}, \gamma_i)), \tag{16}$$

where I is an indicator function. The expression in (16) estimates the p-value as the proportion of the d replications in which the simulated discrepancy variable exceeds the realized value. A p-value close to zero or one, (i.e., $\mid p - 0.5 \mid$ close to 0.5) indicates that the model is inadequate for the aspects measured by the discrepancy variable T.

In binary factor analysis, we suggest test statistics based on coefficients of correlations for 2×2 contingency tables of manifest variables. Tetrachoric correlations can be used for these purposes. Bartholomew (1987, pp. 115–120) discusses other alternatives to the tetrachoric correlation that are easier to compute. If the cross product ratio is given by $\theta = \frac{ad}{bc}$ where a and d are the diagonal frequencies and b and c are the off-diagonal frequencies in the contingency table, then Chambers (1982) shows that a correlation coefficient based on C-type distributions (Mardia, 1970) can be approximated by

$$T(y_i) = r = \frac{\theta^{\nu} - 1}{\theta^{\nu} + 1}, \tag{17}$$

where $\nu = 0.74$, leads to a good approximation. If a model consistently overpredicts or underpredicts a correlation then we can conclude that the correlation structure implied by the model fails in replicating that correlation in the actual data. In the simulations to follow, we illustrate the diagnostic potential of such correlation coefficients.

In situations where the binary variables p are limited in number, the frequency distributions of the 2^p different response profiles[2] can be compared across the actual and the replicated data sets. Such a comparison can yield comprehensive information about the adequacy of the model in capturing marginal probabilities of various orders. In addition to the above test statistics, we can also use discrepancy variables based on Bayesian residuals described in Albert and Chib (1995). The model $\mathbf{u}_{ij} = \mathbf{m}_i + \mathbf{\Lambda}_1 \delta_{1,ij} + \boldsymbol{\varepsilon}_{ij}$ suggests the Bayesian latent residuals $\boldsymbol{\varepsilon}_{ijk}(\mathbf{m}_{ik}, \mathbf{\Lambda}_1, \boldsymbol{\delta}_{1,ij}) = u_{ijk} - (m_{ik} + \boldsymbol{\lambda}_{1,k}' \boldsymbol{\delta}_{1,ij})$, for $k = 1$ to p. These latent residuals are available easily as a by-product of the MCMC simulation. Various summary measures of these residuals can be utilized to assess model adequacy. For example, $Q - Q$ plots can be utilized to test the normality assumptions of the measurement errors.

4.2. *Model Comparison*

Bayes factors (Kass & Raftery, 1995) have traditionally been used in Bayesian analysis to compare two models. Chib and Greenberg (1998) discuss the computation of the Bayes factor for multivariate probit models. Their computational approach is difficult to use in our multilevel factor analysis setting owing to the identifiability constraints that are necessary at the first level of our model. We therefore use the pseudo-Bayes factor (PsBF) (Gelfand, 1996; Sahu, 1998) as a surrogate for the Bayes factor. The PsBF is based on the cross-validation predictive density of the data instead of the prior predictive density used in the calculation of Bayes factors. It can therefore be used even with improper priors. Moreover, it can be very conveniently computed using the MCMC draws for our model.

[2]A response profile is given by a sequence of ones and zeros on the p binary variables.

Let $\mathbf{y}$ be the observed data and let $\mathbf{y}_{(ijk)}$ represent the data with the kth variable of observation j from group i deleted. The cross-validation predictive density can then be written as

$$\pi(y_{ijk} \mid \mathbf{y}_{(ijk)}) = \int \pi(y_{ijk} \mid \boldsymbol{\gamma}, \mathbf{y}_{(ijk)})\pi(\boldsymbol{\gamma} \mid \mathbf{y}_{(ijk)})d\boldsymbol{\gamma} \tag{18}$$

where γ is the vector of all parameters in the model. The PsBF for comparing two models (M1 and M2) is expressed in terms of the product of cross-validation predictive densities and can be written as

$$\text{PsBF} = \prod_{i=1}^{I}\prod_{j=1}^{n_i}\prod_{k=1}^{p}\frac{\pi(y_{ijk} \mid \mathbf{y}_{(ijk)}, M1)}{\pi(y_{ijk} \mid \mathbf{y}_{(ijk)}, M2)}. \tag{19}$$

The PsBF summarizes the evidence provided by the data for M1 against M2 and its value can be interpreted as the number of times model M1 is more (or less) probable than model M2.

The PsBF for our model can be calculated easily from a sample of d MCMC draws $\{\gamma_1, \ldots, \gamma_d\}$. As γ is the vector of all parameters, including the factor scores, the binary responses y_{ijk}, $i = 1$ to I, $j = 1$ to n_i and $k = 1$ to p, are conditionally independent given $\boldsymbol{\gamma}$. In such a situation, a Monte Carlo esimate of $\pi(y_{ijk} \mid \mathbf{y}_{(ijk)})$ can be obtained as

$$\hat{\pi}(y_{ijk} \mid \mathbf{y}_{(ijk)}) = \left(\frac{1}{d}\sum_{t=1}^{d}\frac{1}{(p_{ijk}^{(t)})^{y_{ijk}}(1-p_{ijk}^{(t)})^{1-y_{ijk}}}\right)^{-1}. \tag{20}$$

In (20), $p_{ijk}^{(t)}$ is the probability $\Pr(y_{ijk} = 1 \mid \boldsymbol{\gamma}_t) = 1 - \Phi\left(\frac{u_{ijk}^{(t)} - m_{ijk}^{(t)} - (\boldsymbol{\lambda}_{1,k}^{(t)})'\boldsymbol{\delta}_{1ij}^{(t)}}{\sqrt{(\theta_{1,kk}^{(t)})}}\right)$, where the superscript t denotes the t-th draw. Gelfand (1996) provides the derivation for Equation (20). In practice, we can calculate the logarithms of the numerator and denominator of the PsBF and these can be used for comparing different models.

5. Simulations

We investigate the MCMC procedures described above using three simulation studies. The first simulation examines the performance of the algorithms in recovering the true parameters. The second simulation examines the performance of the different criteria for model assessment. The third simulation examines the sensitivity of the parameter estimates to different hyperparameter specifications.

5.1. First Simulation: Parameter Recovery

To assess how well the MCMC procedures recover the true simulated parameters, we used a balanced 8 variate data set with 125 groups and 30 observations within each group according to the model in (5) and (7). We set $\boldsymbol{\mu} = 0$,

$$\boldsymbol{\Lambda}_1' = \boldsymbol{\Lambda}_2' = \begin{pmatrix} 0.9 & 0.96 & 0.9 & 0.9 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.9 & 0.92 & 0.96 & 0.9 \end{pmatrix}, \tag{21}$$

$$\boldsymbol{\Psi}_1 = \boldsymbol{\Psi}_2 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}, \tag{22}$$

$\boldsymbol{\Theta}_1 = \mathbf{I} - \text{diag}(\boldsymbol{\Lambda}_1\boldsymbol{\Psi}_1\boldsymbol{\Lambda}_1)$ and $\boldsymbol{\Theta}_2 = 0.1\mathbf{I}$, to generate the simulated data.

We used priors that are similar to those outlined in section 2.2. The prior for $\boldsymbol{\mu}$ is assumed to be $p(\boldsymbol{\mu}) = N(0, 100I)$. The second level variances have independent inverse gamma priors each given by $p(\theta_{kk}) = IG(0.001, 1000)$, $k = 1, \ldots p$. The priors for the two correlation matrices are as described in Section 2.2 with $\text{vec}(\boldsymbol{\Psi}_{1,0}) = \text{vec}(\boldsymbol{\Psi}_{2,0}) = \mathbf{0}$ and $\mathbf{G}_{1,0} = \mathbf{G}_{2,0} = I$. Finally

we assume independent univariate normal $N(0, 100)$ priors over the individual elements in $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$.

We used a total of 50 Monte Carlo replications to study the variation across the generated samples. For each of the 50 data sets we estimated the model parameters based on 3500 draws from the joint posterior distribution, after an initial transient phase of 1500 draws. We used a single chain for each estimation to make the computations manageable. We checked convergence using the Geweke's spectral density diagnostic (Geweke, 1992) which is part of the CODA package (Best, Cowles & Vines, 1995) and by using time series plots to graphically assess the quality of the mixing in the chain. In this simulation, the Geweke convergence diagnostic for most parameters was within the ± 1.96 range indicating that convergence is plausible. Computations were performed on a Sun Enterprise 4000 machine using programs written in the C language by the authors. The time for each run of 5000 iterations is approximately 50 minutes.

Table 1 reports the true parameters, the average of the mean and the average of standard deviation of the posterior distributions of the parameters over the 50 Monte Carlo samples. The Table also includes the 95% coverage for each parameter. The coverage is the proportion of the 50 Monte Carlo samples in which the 95% posterior interval spanning the 2.5th to the 97.5th percentile of the MCMC draws covers the true parameters. Table 1 shows close agreement between the true parameter and the average estimated mean across the samples. The coverage properties are also good considering that these are based on as little as 50 Monte Carlo samples.

5.2. *Second Simulation: Model Assessment Criteria*

In order to assess the performance of the model adequacy test quantity described in Section 4.1 and to investigate the performance of the pseudo Bayes factor for model comparison, we estimated two alternative models with factor structures that are different from that of the true model. We then computed the posterior predictive p-values for the correlations in (17) and the PsBF. The first alternative model (Model 2) assumes a *single* factor while the second alternative model (Model 3) assumes two factors as in the true model (Model 1) but with a factor structure given by the loading's matrices

$$\mathbf{\Lambda}_1' = \mathbf{\Lambda}_2' = \begin{pmatrix} \lambda_{11} & \lambda_{21} & \lambda_{31} & 0 & 0 & 0 & 0 & \lambda_{81} \\ 0 & 0 & 0 & \lambda_{42} & \lambda_{52} & \lambda_{62} & \lambda_{72} & 0 \end{pmatrix}. \tag{23}$$

Note that Model 3 is misspecified as indicator y_4 incorrectly loads on factor 2, whereas indicator y_8 incorrectly loads on factor 1.

Table 2 reports the results of the posterior predictive checks associated with the $p(p-1)/2$ nonredundant correlations between the manifest variables for the three models. The models were estimated for each of the 50 samples generated using the true parameters in Section 5.1. The p-values for model adequacy were computed based on 1500 replicated data sets for each original sample. The first set of results in Table 2 gives the mean across the 50 samples of the mean absolute deviations $| r_{act} - r_{rep} |$ for each model and each correlation, where r_{act} is the correlation based on the actual data and r_{rep} is the correlation from a replicated data set. The fourth column of Table 2, (Model 3) shows that the absolute deviations associated with variables 4 and 8 (e.g., r_{14}, r_{24}, r_{58}) are of larger magnitude than those corresponding to the other variables. This indicates that the factors associated with these variables may have been misspecified. The second set of results in Table 2 report the average across the 50 samples of the absolute deviation of the p-values from 0.5, $| p - 0.5 |$, for each model and each correlation. It is clear from Table 2 that the p-values for Model 1 are all near 0.5 resulting in $| p - 0.5 |$ being small for all correlations. This is expected as Model 1 has the true factor structure. The p-values for Model 2, the single factor model, are all extreme, clearly indicating that a single factor does not adequately capture the association structure in the manifest variables. The last column of Table 2 indicates that Model 3 also does not adequately capture the nature of associations in the data. The average pseudo-Bayes factor across the fifty samples for Model 1 versus Model 2 is given by $\exp(-8645.27 +$

TABLE 1.
Simulation 1: Parameter Recovery

Level One Estimates				
Parameter	True	Mean	Std. Dev.	Coverage
λ_{11}	0.9	0.895	0.01	0.96
λ_{21}	0.96	0.959	0.007	0.88
λ_{31}	0.9	0.893	0.01	0.96
λ_{41}	0.9	0.898	0.01	0.88
λ_{52}	0.9	0.902	0.009	0.92
λ_{62}	0.92	0.918	0.009	0.98
λ_{72}	0.96	0.959	0.006	0.88
λ_{82}	0.9	0.898	0.01	0.96
Ψ_{12}	0.5	0.495	0.02	0.96
Level Two Estimates				
μ_1	0	−0.021	0.09	0.98
μ_2	0	−0.02	0.095	1
μ_3	0	−0.013	0.09	1
μ_4	0	−0.029	0.092	0.96
μ_5	0	0.012	0.09	1
μ_6	0	0.022	0.091	0.92
μ_7	0	0.023	0.094	0.96
μ_8	0	0.015	0.089	0.96
λ_{11}	0.9	0.904	0.077	0.92
λ_{21}	0.96	0.97	0.081	0.92
λ_{31}	0.9	0.896	0.076	1
λ_{41}	0.9	0.925	0.078	0.92
λ_{52}	0.9	0.932	0.077	0.88
λ_{62}	0.92	0.939	0.077	0.88
λ_{72}	0.96	0.981	0.079	0.92
λ_{82}	0.9	0.912	0.074	0.88
Ψ_{12}	0.5	0.481	0.078	0.96
θ_{11}	0.1	0.108	0.025	0.96
θ_{22}	0.1	0.102	0.025	0.96
θ_{33}	0.1	0.108	0.025	1
θ_{44}	0.1	0.104	0.025	0.96
θ_{55}	0.1	0.1	0.024	0.88
θ_{66}	0.1	0.102	0.024	0.92
θ_{77}	0.1	0.097	0.024	0.92
θ_{88}	0.1	0.094	0.023	0.96

10963.05) = exp(2317.78) and the average PsBF for Model 1 against Model 3 is given by exp(−8645.27 + 10444.19) = exp(1798.92) clearly indicating strong support for the true model over the two misspecified models.

5.3. *Third Simulation: Hyperparameter Sensitivity*

In this simulation we focussed on further studying the accuracy of the results and the sensitivity of parameter estimates to different specifications of hyperparameter values. We used a balanced 8 variate data set with 200 groups and 20 observations within each group according to the model in (5) and (7). We set $\boldsymbol{\mu} = 0$,

TABLE 2.
Simulation 2: Posterior Predictive Checking

	Mean $\mid r_{act} - r_{rep} \mid$			Mean $\mid p - 0.5 \mid$		
Correlation	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
r_{12}	0.009	0.154	0.025	0.106	0.5	0.47
r_{13}	0.011	0.146	0.023	0.137	0.5	0.429
r_{14}	0.01	0.147	0.205	0.132	0.5	0.5
r_{15}	0.018	0.108	0.076	0.177	0.5	0.5
r_{16}	0.017	0.112	0.08	0.167	0.5	0.5
r_{17}	0.016	0.112	0.078	0.163	0.5	0.5
r_{18}	0.018	0.111	0.092	0.185	0.5	0.49
r_{23}	0.009	0.155	0.026	0.099	0.5	0.47
r_{24}	0.009	0.153	0.216	0.125	0.5	0.5
r_{25}	0.016	0.112	0.078	0.18	0.5	0.5
r_{26}	0.014	0.114	0.08	0.106	0.5	0.5
r_{27}	0.014	0.116	0.078	0.132	0.5	0.5
r_{28}	0.015	0.113	0.096	0.135	0.5	0.5
r_{34}	0.01	0.149	0.207	0.123	0.5	0.5
r_{35}	0.016	0.11	0.079	0.142	0.5	0.5
r_{36}	0.016	0.113	0.081	0.142	0.5	0.5
r_{37}	0.017	0.117	0.083	0.189	0.5	0.5
r_{38}	0.017	0.108	0.09	0.143	0.5	0.5
r_{45}	0.017	0.109	0.087	0.192	0.5	0.5
r_{46}	0.016	0.113	0.092	0.144	0.5	0.5
r_{47}	0.016	0.116	0.095	0.178	0.5	0.5
r_{48}	0.017	0.11	0.03	0.168	0.5	0.34
r_{56}	0.009	0.07	0.018	0.121	0.5	0.42
r_{57}	0.008	0.073	0.021	0.084	0.5	0.45
r_{58}	0.01	0.068	0.203	0.122	0.5	0.5
r_{67}	0.008	0.076	0.021	0.122	0.5	0.46
r_{68}	0.01	0.07	0.206	0.109	0.5	0.5
r_{78}	0.009	0.075	0.216	0.138	0.5	0.5

$$\mathbf{\Lambda}_1' = \begin{pmatrix} 0.8 & 0.8 & 0.8 & 0.8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.75 & 0.8 & 0.8 & 0.75 \end{pmatrix}, \tag{24}$$

$$\mathbf{\Lambda}_2' = \begin{pmatrix} 0.75 & 0.8 & 0.8 & 0.75 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.8 & 0.8 & 0.8 & 0.8 \end{pmatrix}, \tag{25}$$

$$\mathbf{\Psi}_1 = \mathbf{\Psi}_2 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}, \tag{26}$$

$\mathbf{\Theta}_1 = \mathbf{I} - \text{diag}(\mathbf{\Lambda}_1\mathbf{\Psi}_1\mathbf{\Lambda}_1)$ and $\mathbf{\Theta}_2 = 0.2\mathbf{I}$, to generate the simulated data.

We used a total of 50 Monte Carlo samples from the above model. For each of these 50 samples we estimated the model using two sets of priors. In the first set, the prior for $\boldsymbol{\mu}$ is assumed to be $p(\boldsymbol{\mu}) = N(0, 100I)$. The second level variances have independent inverse gamma priors each given by $p(\theta_{kk}) = IG(0.001, 1000), k = 1, \ldots p$. The priors for the two correlation matrices are as described in Section 2.2 with $\text{vec}(\mathbf{\Psi}_{1,0}) = \text{vec}(\mathbf{\Psi}_{2,0}) = \mathbf{0}$ and $\mathbf{G}_{1,0} = \mathbf{G}_{2,0} = I$. Finally we assume independent univariate normal $N(0, 100)$ priors over the individual elements in $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$. In the second set, the prior for $\boldsymbol{\mu}$ is assumed to be $p(\boldsymbol{\mu}) = N(0, 1000I)$. The second level variances are assumed to be independent inverse gamma priors each given by $p(\theta_{kk}) = IG(3, 1000), k = 1, \ldots p$. The priors for the two correlation matrices are as described

in Section 2.2 with $\text{vec}(\boldsymbol{\Psi}_{1,0}) = \text{vec}(\boldsymbol{\Psi}_{2,0}) = \mathbf{0}$ and $\mathbf{G}_{1,0} = \mathbf{G}_{2,0} = 0.5I$. Finally we assume independent univariate normal $N(0, 1000)$ priors over the individual elements in $\boldsymbol{\Lambda}_1$ and $\boldsymbol{\Lambda}_2$.

Table 3 reports the average results from the fifty simulated data sets for the two sets of priors. These results are again based on 3500 iterations after an initial burn in period of 1500 iterations. The Geweke Convergence diagnostic for most parameters was within the ± 1.96 range indicating that convergence is plausible. The time series plots of major parameters also indicated good mixing of the chain. As is clear from the table, the parameter estimates are virtually identical in value and appear insensitive to the specification of hyperparameter values as these are sufficiently diffuse in nature in both sets of specifications.

TABLE 3.
Simulation 3: Hyper-Parameter Sensitivity

Level One Estimates							
		Prior 1			Prior 2		
Parameter	True	Mean	Std. Dev	Coverage	Mean	Std. Dev	Coverage
λ_{11}	0.8	0.797	0.017	1	0.797	0.017	1
λ_{21}	0.8	0.796	0.017	0.9	0.796	0.017	0.9
λ_{31}	0.8	0.797	0.017	0.98	0.796	0.017	0.94
λ_{41}	0.8	0.8	0.017	0.98	0.8	0.017	0.98
λ_{52}	0.75	0.746	0.019	0.96	0.746	0.019	0.94
λ_{62}	0.8	0.802	0.018	0.88	0.802	0.018	0.88
λ_{72}	0.8	0.802	0.017	0.98	0.801	0.018	0.94
λ_{82}	0.75	0.75	0.018	0.88	0.75	0.018	0.9
Ψ_{12}	0.5	0.499	0.022	0.92	0.499	0.022	0.94
Level two estimates							
μ_1	0	−0.006	0.066	0.98	−0.005	0.066	0.98
μ_2	0	−0.001	0.068	0.94	0.001	0.069	0.96
μ_3	0	0.007	0.069	0.96	0.008	0.069	0.96
μ_4	0	0.009	0.066	0.96	0.01	0.067	0.96
μ_5	0	0.001	0.069	0.94	0.002	0.069	0.98
μ_6	0	0.013	0.068	0.94	0.014	0.069	0.94
μ_7	0	0.021	0.068	0.94	0.022	0.07	0.94
μ_8	0	0.02	0.068	0.9	0.022	0.069	0.94
λ_{11}	0.75	0.768	0.062	0.88	0.767	0.061	0.88
λ_{21}	0.8	0.823	0.064	0.84	0.818	0.063	0.86
λ_{31}	0.8	0.817	0.064	0.92	0.815	0.063	0.92
λ_{41}	0.75	0.767	0.062	0.92	0.767	0.061	0.94
λ_{52}	0.8	0.815	0.063	0.94	0.816	0.063	0.96
λ_{62}	0.8	0.798	0.063	0.92	0.801	0.062	0.88
λ_{72}	0.8	0.81	0.063	0.92	0.812	0.063	0.94
λ_{82}	0.8	0.811	0.063	0.88	0.811	0.063	0.92
Ψ_{12}	0.5	0.498	0.063	0.96	0.497	0.064	0.96
θ_{11}	0.2	0.188	0.035	0.98	0.186	0.035	0.96
θ_{22}	0.2	0.179	0.037	0.94	0.182	0.036	0.96
θ_{33}	0.2	0.187	0.037	0.86	0.186	0.037	0.88
θ_{44}	0.2	0.196	0.036	0.94	0.195	0.036	0.92
θ_{55}	0.2	0.187	0.037	0.9	0.188	0.038	0.9
θ_{66}	0.2	0.193	0.036	0.86	0.195	0.037	0.86
θ_{77}	0.2	0.187	0.036	0.94	0.186	0.037	0.96
θ_{88}	0.2	0.185	0.038	0.9	0.186	0.037	0.92

In summary, our simulation results indicate that our MCMC algorithm does well in recovering the true parameters. In addition, our model adequacy test function and the pseudo Bayes factor measure are effective in model diagnosis and selection. Finally, our procedure is robust to different hyper-parameter values.

6. Application: Second International Mathematics Study

We illustrate our procedures using the mathematics achievement data on U.S. eighth-grade students from the Second International Mathematics Study (Crosswhite, Dossey, Swafford, Mcknight, & Cooney, 1985). Longford and Muthén (1992) and Muthén (1994) analyzed the data to estimate a two-level model involving continuous variables. In our application, we assume a two-level structure with 274 classes as the second level units and 5601 students nested within these classes as the first level units. The average number of students in a classroom is 20. The objective of our illustration is to model the covariation in student achievement on questions belonging to different areas of mathematics. The data set we use has students test results on 12 items, 3 items each from the areas of arithmetic, algebra, measurement and geometry from the post-test questionnaire.

The covariation in the student responses across the items can be modeled in terms of underlying ability factors. Previous research using the SIMS data has investigated the possibility of a single underlying factor pertaining to general mathematics ability. Muthén (1994) estimates a model with a single factor at both levels. As Longford and Muthén (1992) point out, there is considerable heterogeneity in the mathematics curriculum across classrooms due to tracking into enriched, remedial, typical and algebra classes. The opportunity to develop abilities in the four areas of mathematics differs across the classes as the composition of the topics that are covered and the emphasis that is placed on different topics varies across the tracks. For example, typical classes emphasize arithmetic whereas more advanced classes teach algebra and geometry earlier than in typical classes. Such heterogeneity needs to be explicitly modeled in order to draw proper inferences from the data.

Considering the composition of the items involved in our study, the covariation in test results could also be explained in terms of four underlying factors pertaining to abilities in the four areas of arithmetic, algebra, measurement and geometry. We therefore estimate two models on the data. The first model is a single factor model at both levels, whereas the second model assumes a four-factor structure at both levels of the hierarchy. Student's abilities in the four areas are likely to covary, therefore, we allow the four factors to be correlated. As we expect the covariation in the mean achievement at the classroom level to reflect the covariation in responses at the student level, we choose the same factor structure at both levels. The loadings matrices at both levels, therefore, have the same structure but the magnitudes of the free elements are allowed to vary.

We apply the MCMC procedures developed in section 3 on the data to estimate both models. The prior for $\boldsymbol{\mu}$ is assumed to be $p(\boldsymbol{\mu}) = N(0, 100I)$. The second level variances have independent inverse gamma priors each given by $p(\theta_{kk}) = IG(0.001, 1000)$, $k = 1, \ldots p$. The priors for the two correlation matrices for our four factor model are as described in Section 2.2 with $\text{vec}(\boldsymbol{\Psi}_{1,0}) = \text{vec}(\boldsymbol{\Psi}_{2,0}) = \mathbf{0}$ and $\mathbf{G}_{1,0} = \mathbf{G}_{2,0} = I$. Finally we assume independent univariate normal $N(0, 100)$ priors over the individual elements in $\boldsymbol{\Lambda}_1$ and $\boldsymbol{\Lambda}_2$. We ran two chains from very different starting values for 10,000 iterations and monitored convergence using the interval based potential scale reduction factor (PsBF) suggested in Brooks and Gelman (1998). The PsBF was close to one for all parameters and the results reported for the application are based on 15,000 draws from the combined output of both chains after discarding the initial 2500 draws from each chain.

We first compare the adequacy of the two models by focusing on the diagnostics. Table 4 reports the posterior predictive checks associated with the 66 correlations between the 12 items. Recall that a model fails in capturing a particular correlation if the p-value is extreme, that is, $| p - 0.5 |$ is close to 0.5. The p-values were estimated from 1000 replicated data sets generated as

TABLE 4.
Application: SIMS Achievement Study

Posterior Predictive p-values for Correlations $\mid p - 0.5 \mid$					
Correlation	Model 1	Model 2	Correlation	Model 1	Model 2
$r_{1,2}$	0.49	0.061	$r_{4,8}$	0.221	0.028
$r_{1,3}$	0.481	0.049	$r_{4,9}$	0.242	0.362
$r_{1,4}$	0.279	0.189	$r_{4,10}$	0.405	0.062
$r_{1,5}$	0.183	0.101	$r_{4,11}$	0.135	0.048
$r_{1,6}$	0.195	0.222	$r_{4,12}$	0.156	0.335
$r_{1,7}$	0.175	0.393	$r_{5,6}$	0.463	0.139
$r_{1,8}$	0.288	0.074	$r_{5,7}$	0.197	0.025
$r_{1,9}$	0.063	0.252	$r_{5,8}$	0.076	0.125
$r_{1,10}$	0.133	0.461	$r_{5,9}$	0.322	0.394
$r_{1,11}$	0.424	0.208	$r_{5,10}$	0.301	0.128
$r_{1,12}$	0.447	0.243	$r_{5,11}$	0.103	0.271
$r_{2,3}$	0.475	0.069	$r_{5,12}$	0.046	0.161
$r_{2,4}$	0.213	0.264	$r_{6,7}$	0.455	0.437
$r_{2,5}$	0.255	0.206	$r_{6,8}$	0.345	0.401
$r_{2,6}$	0.127	0.112	$r_{6,9}$	0.388	0.367
$r_{2,7}$	0.307	0.010	$r_{6,10}$	0.161	0.238
$r_{2,8}$	0.012	0.310	$r_{6,11}$	0.248	0.150
$r_{2,9}$	0.164	0.135	$r_{6,12}$	0.163	0.013
$r_{2,10}$	0.203	0.307	$r_{7,8}$	0.491	0.102
$r_{2,11}$	0.443	0.229	$r_{7,9}$	0.5	0.099
$r_{2,12}$	0.396	0.112	$r_{7,10}$	0.435	0.492
$r_{3,4}$	0.409	0.438	$r_{7,11}$	0.113	0.013
$r_{3,5}$	0.205	0.246	$r_{7,12}$	0.243	0.119
$r_{3,6}$	0.409	0.445	$r_{8,9}$	0.489	0.057
$r_{3,7}$	0.152	0.226	$r_{8,10}$	0.038	0.341
$r_{3,8}$	0.281	0.146	$r_{8,11}$	0.011	0.177
$r_{3,9}$	0.499	0.484	$r_{8,12}$	0.211	0.023
$r_{3,10}$	0.147	0.472	$r_{9,10}$	0.473	0.339
$r_{3,11}$	0.401	0.162	$r_{9,11}$	0.089	0.139
$r_{3,12}$	0.159	0.234	$r_{9,12}$	0.113	0.012
$r_{4,5}$	0.4	0.242	$r_{10,11}$	0.392	0.294
$r_{4,6}$	0.315	0.074	$r_{10,12}$	0.407	0.288
$r_{4,7}$	0.063	0.141	$r_{11,12}$	0.5	0.295

part of the overall MCMC simulation. It is clear from Table 4 that the four-factor model (Model 2) does a better job at recovering the correlations between the variables. The replicated data generated from the four-factor model appears to be in synchrony with the actual data for most correlations. The pseudo Bayes factor (PsBF) for comparing the four-factor model against the one-factor model is $\exp(-36550.412 + 36633.497) = \exp(83.085)$ and provides clear support for the four factor model.

The level-one posterior means and the standard deviations for the entries in $\mathbf{\Lambda}_1$ an $\mathbf{\Psi}_1$ are given in Table 5 for the four factor model. Table 6 reports the level-two parameter estimates. The level-two factor loadings are positive indicating that it is important to account for across classroom heterogeneity. In addition, the level-two factors are positively correlated. This indicates that there may be unobserved factors that are common to the four factors that we have specificed. Similarly, the factors are strongly correlated at the student level, suggesting a second order factor that may further explain the covariation in the factor scores. Figure 1 shows the level-two factor scores for classrooms, whereas Table 7 shows point estimates and standard deviations of the fac-

TABLE 5.
Application: SIMS Achievement Study

Level-One Parameter Estimates		
Parameter	Estimates	Std. Dev.
$\lambda_{1,1}$	0.508	0.031
$\lambda_{2,1}$	0.476	0.031
$\lambda_{3,1}$	0.529	0.034
$\lambda_{4,2}$	0.364	0.031
$\lambda_{5,2}$	0.420	0.034
$\lambda_{6,2}$	0.237	0.033
$\lambda_{7,3}$	0.511	0.041
$\lambda_{8,3}$	0.357	0.041
$\lambda_{9,3}$	0.560	0.057
$\lambda_{10,4}$	0.367	0.031
$\lambda_{11,4}$	0.507	0.036
$\lambda_{12,4}$	0.554	0.039
Ψ_{12}	0.816	0.057
Ψ_{13}	0.648	0.040
Ψ_{14}	0.653	0.043
Ψ_{23}	0.684	0.073
Ψ_{24}	0.738	0.056
Ψ_{34}	0.623	0.047

tor scores for selected students and classrooms. These factor scores can be used to separate out worse performing students and classrooms for remedial action.

7. Conclusions

We develop procedures for performing simulation based Bayesian inference and model assessment for multilevel binary factor analysis. The procedures developed in the paper circumvent the need for complex multidimensional integration which is necessary for maximum likelihood solutions. Our analysis of simulated data indicates that the MCMC procedure does a good job in recovering the true parameters of the model. The posterior predictive checking procedures are diagnostic in revealing a lack of fit of the wrong models. Although our procedures were presented in the context of binary data, they can accommodate metric and mixed (metric or binary) data situations as special cases. We concentrated on confirmatory factor analysis in the paper, but the procedures can also be used for exploratory factor analysis models. The MCMC approach developed in the paper uses data augmentation and therefore enables the simultaneous estimation of factor scores at all levels of a multilevel hierarchy. Our algorithms can also be naturally extended to data structures with multiple levels of nesting and can also be modified easily to include regressors at all levels of the hierarchy. The hierarchical Bayesian approach allows a seamless transition to higher level models and the MCMC simulation procedures require a few additional steps from the relevant full conditional distributions. The Bayesian approach also has promise for estimating more complex data structures and for handling more general multilevel covariance structure models. Further work is required to develop MCMC algorithms for such general models.

TABLE 6.
Application: SIMS Achievement Study

Level-Two Parameter Estimates		
Parameter	Estimates	Std. Dev.
$\lambda_{1,1}$	0.508	0.029
$\lambda_{2,1}$	0.569	0.031
$\lambda_{3,1}$	0.607	0.034
$\lambda_{4,2}$	0.608	0.033
$\lambda_{5,2}$	0.596	0.000
$\lambda_{6,2}$	0.553	0.043
$\lambda_{7,3}$	0.575	0.031
$\lambda_{8,3}$	0.397	0.027
$\lambda_{9,3}$	0.625	0.039
$\lambda_{10,4}$	0.342	0.028
$\lambda_{11,4}$	0.651	0.041
$\lambda_{12,4}$	0.792	0.045
Ψ_{12}	0.888	0.020
Ψ_{13}	0.883	0.023
Ψ_{14}	0.817	0.030
Ψ_{23}	0.895	0.021
Ψ_{24}	0.773	0.032
Ψ_{34}	0.850	0.030
μ_1	0.100	0.019
μ_2	0.048	0.019
μ_3	−0.153	0.020
μ_4	0.006	0.019
μ_5	0.211	0.019
μ_6	−0.928	0.025
μ_7	−0.059	0.019
μ_8	0.126	0.017
μ_9	1.378	0.027
μ_{10}	−0.436	0.018
μ_{11}	0.172	0.021
μ_{12}	0.196	0.022
$\theta_{1,1}$	0.029	0.010
$\theta_{2,2}$	0.019	0.011
$\theta_{3,3}$	0.063	0.016
$\theta_{4,4}$	0.067	0.016
$\theta_{5,5}$	0.001	0.003
$\theta_{6,6}$	0.234	0.037
$\theta_{7,7}$	0.026	0.013
$\theta_{8,8}$	0.028	0.009
$\theta_{9,9}$	0.033	0.016
$\theta_{10,10}$	0.042	0.012
$\theta_{11,11}$	0.107	0.023
$\theta_{12,12}$	0.074	0.027

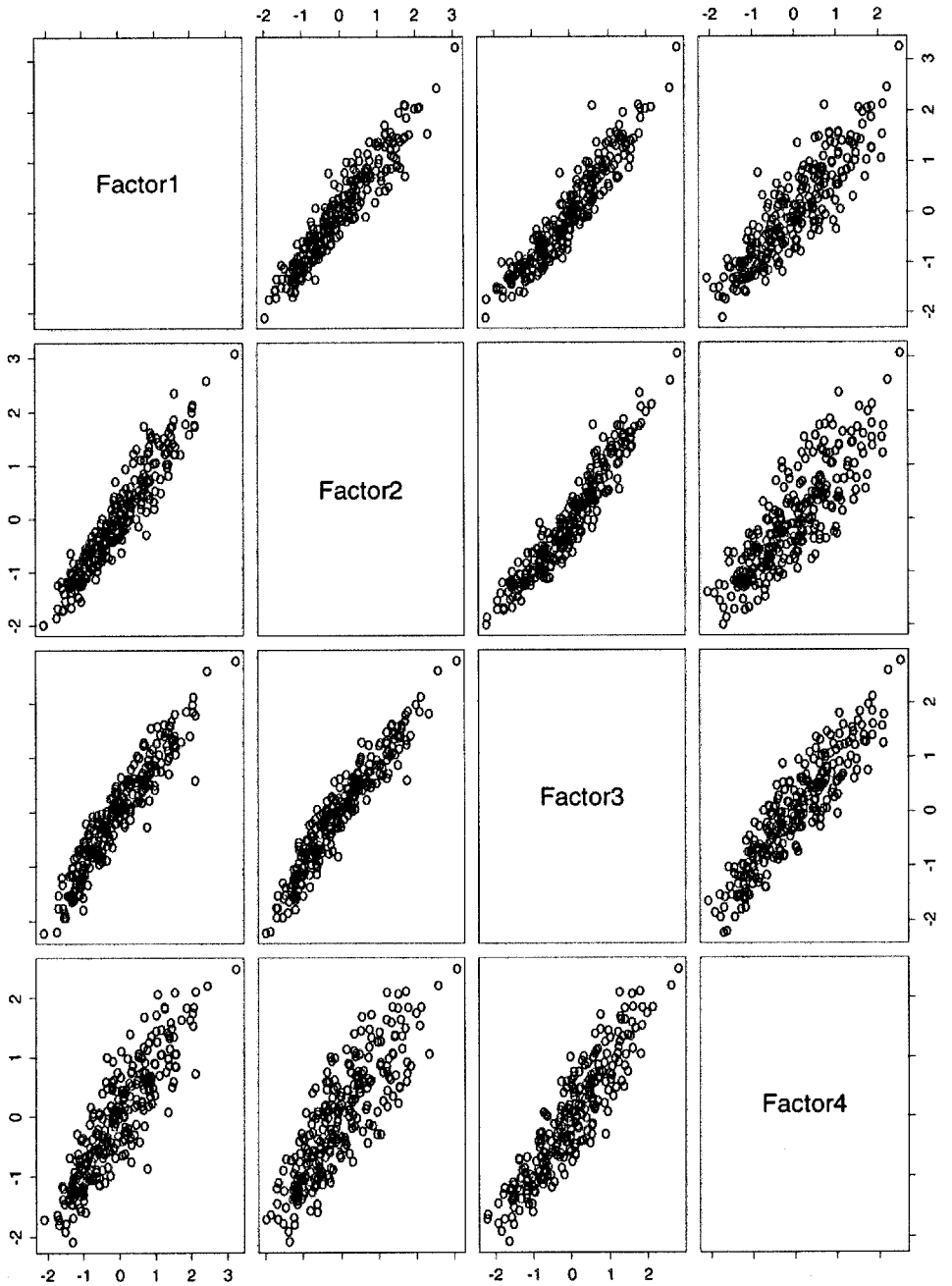

FIGURE 1.
Classroom level factor scores.

TABLE 7.
Application: SIMS Achievement Study

Student Factor Scores: Mean and Standard Deviation				
Obs.	Factor 1	Factor 2	Factor 3	Factor 4
1	−0.612 (0.714)	−0.655 (0.732)	−0.950 (0.732)	−0.76499 (0.733)
2	0.596 (0.729)	0.449 (0.747)	0.782 (0.788)	0.692 (0.777)
⋮	⋮	⋮	⋮	⋮
5600	−0.498 (0.717)	−0.764 (0.735)	−0.550 (0.763)	−0.692 (0.749)
5601	−0.213 (0.719)	−0.143 (0.746)	0.226 (0.786)	0.071 (0.769)
Classroom Factor Scores: Mean and Standard Deviation				
Obs.	Factor 1	Factor 2	Factor 3	Factor 4
1	−0.361 (0.292)	−0.382 (0.292)	−0.151 (0.301)	0.558 (0.342)
2	2.021 (0.349)	2.069 (0.357)	2.062 (0.362)	1.826 (0.387)
⋮	⋮	⋮	⋮	⋮
273	1.538 (0.316)	1.297 (0.308)	1.146 (0.319)	1.045 (0.336)
274	0.312 (0.349)	0.415 (0.348)	0.485 (0.360)	0.500 (0.378)

A. Appendix: Full Conditional Distributions

The $(m+1)$-th iteration of the substitution sampling algorithm involves generating random draws from the following full conditional distributions:

1. The lower level correlation matrix $\boldsymbol{\Psi}_1$ can be drawn using a Metropolis Hit and Run algorithm (Chen & Schmeiser 1993; Dey & Chen, 1998). An alternative procedure for sampling correlation matrices is described in Chib & Greenberg (1998). Their method can be more efficient for large correlation matrices, however, it requires optimization steps and adjustable parameters for adaptive tuning of the proposal density. We use the hit and run algorithm for its simplicity. Further research is needed to compare the efficiency of Chib and Greenberg procedure with that of the hit and run algorithm. If the prior distribution for the nonredundant and free elements of $\boldsymbol{\Psi}_1$ that are contained in the vector $\text{vec}(\boldsymbol{\Psi}_1)$ is given by $\pi(\text{vec}(\boldsymbol{\Psi}_1) \mid \boldsymbol{\psi}_{1,0}, \mathbf{G}_{1,0})$, as shown in equation (10), then the full conditional of $\boldsymbol{\Psi}_1$ is proportional to the product of the likelihood $L(\boldsymbol{\Psi}_1 \mid \{\{\mathbf{u}_{ij}\}\}, \boldsymbol{\Lambda}_1, \{\mathbf{m}_i\}, \{\{\boldsymbol{\delta}_{1,ij}\}\}, \boldsymbol{\Theta}_1)$ and the prior $\pi(\text{vec}(\boldsymbol{\Psi}_1))$. Here $L(\cdot)$ is the conditional likelihood of observing the "data" $\{\{\mathbf{u}_{ij}\}\}$ given the matrix $\boldsymbol{\Psi}_1$ and the other parameters, and is proportional to

$$|\boldsymbol{\Theta}_1|^{-\frac{N}{2}} \exp\left\{-\tfrac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(\mathbf{u}_{ij}-\mathbf{m}_i-\boldsymbol{\Lambda}_1\boldsymbol{\delta}_{1,ij})'\boldsymbol{\Theta}_1^{-1}(\mathbf{u}_{ij}-\mathbf{m}_i-\boldsymbol{\Lambda}_1\boldsymbol{\delta}_{1,ij})\right\}. \tag{A1}$$

Direct methods for sampling from this full conditional distribution are not available so we generate $\boldsymbol{\Psi}_1$ using a Metropolis Hit-and-Run algorithm. If $\boldsymbol{\Psi}_1^{(m)}$ is the current value of the correlation matrix, then in the $(m+1)^{th}$ step, a candidate matrix $\boldsymbol{\Psi}_1^c$ is generated by specifying a random walk chain $\boldsymbol{\Psi}_1^c = \boldsymbol{\Psi}_1^{(m)} + H$, where $H = (h_{ij})$ is an increment matrix with $E(h_{ij}) = 0$ and $h_{ii} = 0$, for all i and j. Let ξ be the smallest eigenvalue of $\boldsymbol{\Psi}_1^{(m)}$. Then the elements of the increment matrix H can be generated using the Hit-and-Run algorithm which involves the following steps:

(a) generate a sequence of iid standard normal deviates $z_{12}, z_{13}, \ldots, z_{(r_1-1),r_1}$, of length $r_1(r_1-1)/2$

(b) generate a deviate d from $N(0, \sigma_d^2)$ which is truncated to the interval $\left(-\frac{\xi}{\sqrt{2}}, \frac{\xi}{\sqrt{2}}\right)$

(c) formulate the elements

$$h_{ij} = \frac{dz_{ij}}{(\sum_{j=1}^{J-1} \sum_{l=j+1}^{J} z_{jl}^2)^{(1/2)}}$$

for $i < j$, $h_{ii} = 0$, and $h_{ij} = h_{ji}$ for $i > j$.

Here σ_d^2 is a tuning constant that needs to be chosen such that candidates are not rejected disproportionately. If ξ^c is the smallest eigenvalue of the candidate matrix, then once a candidate is generated, it is accepted or rejected based on the following Metropolis–Hastings acceptance probability

$$\min \left\{ \frac{L(\boldsymbol{\Psi}_1^c \mid .)p(\text{vec}(\boldsymbol{\Psi}_1^c)) \left(\Phi(\frac{\xi^c}{\sqrt{2}\sigma_d}) - \Phi\left(\frac{-\xi^c}{\sqrt{2}\sigma_d}\right)\right)}{L(\boldsymbol{\Psi}_1^{(m)} \mid .)p(\text{vec}(\boldsymbol{\Psi}_1^{(m)})) \left(\Phi\left(\frac{\xi}{\sqrt{2}\sigma_d}\right) - \Phi\left(\frac{-\xi}{\sqrt{\sigma_d}}\right)\right)}, 1 \right\} \tag{A2}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. If the candidate is accepted then $\boldsymbol{\Psi}_1^{(m+1)} = \boldsymbol{\Psi}_1^{(c)}$, otherwise $\boldsymbol{\Psi}_1^{(m+1)} = \boldsymbol{\Psi}_1^{(m)}$.

2. The full conditional distribution for the underlying variables $\mathbf{u}_{ij}$ on any observation is a product of p truncated univariate normal distributions. In forming the full conditional, we can utilize the conditional independence property of latent variable models. The variables u_{ijk}, $k = 1$ to p, on each observation are conditionally independent of each other given the lower level factor scores $\boldsymbol{\delta}_{1,ij}$ for that observation. For every observation in the sample, we can therefore sample each of the k variables in sequence from truncated univariate normal distributions. Thus

$$p(\mathbf{u}_{ij}^{(m+1)} \mid \mathbf{y}_{ij}, \mathbf{m}_i, \delta_{1,ij}, \boldsymbol{\Lambda}_1) = \prod_{k=1}^{p} I(s_{ijk}) N(m_{ik} + \boldsymbol{\lambda}_{1k}' \boldsymbol{\delta}_{1,ij}, \theta_{1,kk}) \tag{A3}$$

where $s_{ijk} = (-\infty, 0)$ if $y_{ijk} = 0$, $s_{ijk} = (0, \infty)$, if $y_{ijk} = 1$, and $I(\cdot)$ is an indicator function that determines the support of the truncated normal distribution. In the above distribution, the variance $\theta_{1,kk}$ is the k-th diagonal element of $\boldsymbol{\Theta}_1$ obtained from the deterministic relationship in (6), and $\boldsymbol{\lambda}_{1k}'$ is the k-th row of $\boldsymbol{\Lambda}_1$.

3. The full conditional for the level one factor scores $\boldsymbol{\delta}_{1,ij}$ for observation j belonging to group i is a multivariate normal distribution. This posterior distribution can be derived easily using standard Bayesian theory pertaining to linear models. The prior for $\boldsymbol{\delta}_{1,ij}$ is $N(\mathbf{0}, \boldsymbol{\Psi}_1)$. When this is combined with the likelihood of observing $\mathbf{u}_{ij}$ given $\boldsymbol{\delta}_{1,ij}$, we obtain a multivariate normal posterior full conditional distribution

$$p(\boldsymbol{\delta}_{1,ij} \mid \mathbf{u}_{ij}, \mathbf{m}_i, \boldsymbol{\Lambda}_1, \boldsymbol{\Psi}_1, \boldsymbol{\Theta}_1) = N(\hat{\boldsymbol{\delta}}_{1,ij}, \mathbf{V}_{\delta_1}) \tag{A4}$$

where $\mathbf{V}_{\delta_1}^{-1} = \boldsymbol{\Psi}_1^{-1} + \boldsymbol{\Lambda}_1' \boldsymbol{\Theta}_1^{-1} \boldsymbol{\Lambda}_1$ and $\hat{\boldsymbol{\delta}}_{1,ij} = \mathbf{V}_{\delta_1} \boldsymbol{\Lambda}_1' \boldsymbol{\Theta}_1^{-1} (\mathbf{u}_{ij} - \mathbf{m}_i)$. The factor scores, $\boldsymbol{\delta}_{1,ij}$ are independently sampled for each observation in the sample, i.e., for $i = 1$ to I, $j = 1$ to n_i.

4. The full conditional for the level two factor scores $\boldsymbol{\delta}_{2,i}$ for each group $i = 1$ to I, is given by conditional

$$p(\boldsymbol{\delta}_{2,i} \mid \mathbf{m}_i, \mu, \boldsymbol{\Lambda}_2, \boldsymbol{\Psi}_2, \boldsymbol{\Theta}_2) = N(\hat{\boldsymbol{\delta}}_{2,i}, \mathbf{V}_{\delta_2}) \tag{A5}$$

where $\mathbf{V}_{\boldsymbol{\delta}_2}^{-1} = \boldsymbol{\Psi}_2^{-1} + \boldsymbol{\Lambda}_2' \boldsymbol{\Theta}_2^{-1} \boldsymbol{\Lambda}_2$ and $\hat{\boldsymbol{\delta}}_{2,i} = \mathbf{V}_{\delta_2} \boldsymbol{\Lambda}_2' \boldsymbol{\Theta}_2^{-1} (\mathbf{m}_i - \boldsymbol{\mu})$.

5. The mean vector $\mathbf{m}_i$ for each group, $i = 1$ to I, can be generated from the full conditional distribution

$$p(\mathbf{m}_i \mid \{\mathbf{u}_{ij}\}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\mu}) = N(\hat{\mathbf{m}}_i, \mathbf{V}_{m_i}), \tag{A6}$$

where $\mathbf{V}_{m_i}^{-1} = \boldsymbol{\Sigma}_2^{-1} + n_i\boldsymbol{\Sigma}_1^{-1}$ and $\hat{\mathbf{m}}_i = \mathbf{V}_{m_i}(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu} + \sum_{j=1}^{n_i}\boldsymbol{\Sigma}_1^{-1}\mathbf{u}_{ij})$.

6. The overall mean $\boldsymbol{\mu}$ can be generated from the multivariate normal full conditional distribution given by

$$p(\boldsymbol{\mu} \mid \{\mathbf{m}_i\}, \boldsymbol{\Sigma}_2) = N(\hat{\mu}, \mathbf{V}_\mu), \tag{A7}$$

where $\mathbf{V}_\mu^{-1} = \mathbf{C}^{-1} + I\boldsymbol{\Sigma}_2^{-1}$ and $\hat{\boldsymbol{\mu}} = \mathbf{V}_\mu(\mathbf{C}^{-1}\boldsymbol{\eta} + \sum_{i=1}^{I}\boldsymbol{\Sigma}_2^{-1}\mathbf{m}_i)$

7. The full conditional distributions for the diagonal elements of the matrix $\boldsymbol{\Theta}_2$, that is, $\theta_{2,kk}$, $k = 1$ to p, are independent inverse gamma distributions. These follow from standard Bayesian theory pertaining to linear models. Thus we have

$$p\big(\theta_{2,kk} \mid \boldsymbol{\mu}, \{\mathbf{m}_i\}, \{\boldsymbol{\delta}_{2,i}\}, \boldsymbol{\Lambda}_2\big) = IG\left(\frac{I}{2} + a, \left[\frac{\sum_{i=1}^{I}(m_{ik} - \mu_k - \lambda_k'\boldsymbol{\delta}_{2,i})^2}{2} + b^{-1}\right]^{-1}\right). \tag{A8}$$

8. The full conditional distribution for the nonzero elements within a row of the level-two loadings $\boldsymbol{\Lambda}_2$ is multivariate normal. The full conditionals pertaining to the different rows are independent and therefore the rows can be handled sequentially. The prior for the nonzero elements pertaining to the k-th row is given by $p(\tilde{\boldsymbol{\lambda}}_{2k}) = N(\mathbf{g}_{2k}, \mathbf{H}_2)$. Define the $I \times r_{2k}$ matrix $\mathbf{Z}_{2k}$ containing the level-two factor scores pertaining to the r_{2k} nonzero loadings in the k^{th} row of $\boldsymbol{\Lambda}_2$. Let $\mathbf{vm}_k$ be the I vector containing the group means for the k-th variable. Define a I vector $\mathbf{v\tilde{m}}_k$ whose i-th element is given by $\mathbf{vm}_{ki} - \boldsymbol{\mu}_k$. Given the prior, the vector $\tilde{\boldsymbol{\lambda}}_{2k}$ can be sampled from the full conditional distribution given by

$$p(\tilde{\boldsymbol{\lambda}}_k \mid \mathbf{v\tilde{m}}_k, \mathbf{Z}_{2k}, \theta_{2,kk}) = N(D_k(\theta_{2,kk}^{-1}\mathbf{Z}_{2k}'\mathbf{v\tilde{m}}_k + \mathbf{H}_2^{-1}\mathbf{g}_{2k}), D_k) \tag{A9}$$

where $D_k^{-1} = \theta_{2,kk}^{-1}\mathbf{Z}_{2k}'\mathbf{Z}_{2k} + \mathbf{H}_2^{-1}$.

9. The factor loadings associated with $\boldsymbol{\Lambda}_1$ can be generated one by one from truncated univariate normal distributions. let $\lambda_{1,kl}$ be the element in row k and column l of $\boldsymbol{\Lambda}_1$. Given the prior $p(\lambda_{1,kl}) = N(g_{1,kl}, h_{1,kl})$ and taking into account the constraint that $\theta_{1,kk} = 1 - \lambda_{1k}'\boldsymbol{\Psi}_1\lambda_{1k} > 0$, the posterior full conditional is given by the truncated normal

$$p(\lambda_{1,kl} \mid \{u_{ij}\}, \boldsymbol{\Theta}_1, \delta_1, \boldsymbol{\Psi}_1) = tn(\hat{\lambda}_{1,kl}, v_{\lambda_{1,kl}}), \tag{A10}$$

where $v_{\lambda_{1,kl}}^{-1} = h_{1,kl}^{-1} + \boldsymbol{\delta}_{1,l}'\boldsymbol{\delta}_{1,l}\theta_{1,kk}^{-1}$ and $\hat{\lambda}_{1,kl} = v_{\lambda_{1,kl}}(h_{1,kl}^{-1} + \theta_{1,kk}^{-1}\boldsymbol{\delta}_{1,l}'\tilde{\mathbf{u}}_k)$. The vector $\boldsymbol{\delta}_{1,l}$ contains the level-one factor scores for factor l and $\tilde{\mathbf{u}}_k$ contains the the adjusted threshold values $\tilde{u}_{ijk} = u_{ijk} - m_{ijk} - \boldsymbol{\lambda}_{1,k(-l)}'\boldsymbol{\delta}_{1,ij(-l)}$ where $\boldsymbol{\lambda}_{1,k(-l)}$ is a vector containing the elements from row k of $\boldsymbol{\Lambda}_1$ excluding the kl-th element. The truncation points of the normal distribution can easily be obtained from the constraint $\theta_{1,kk} = 1 - \lambda_{1k}'\boldsymbol{\Psi}_1\lambda_{1k} > 0$. Alternatively, a Metropolis step can be used to obtain the different rows of $\boldsymbol{\Lambda}_1$ and the entire matrix $\boldsymbol{\Lambda}_1$ can be constructed by independently sampling each row in sequence.

10. The level-two correlation matrix $\boldsymbol{\Psi}_2$ can be drawn using a Metropolis Hit and Run algorithm analogous to that in Step 1. If the prior distribution for the nonredundant and free elements of $\boldsymbol{\Psi}_1$ that are contained in the vector vec($\boldsymbol{\Psi}_2$) is given by $\pi(\text{vec}(\boldsymbol{\Psi}_2) \mid \boldsymbol{\psi}_{2,0}, \mathbf{G}_{2,0})$, then the full conditional of $\boldsymbol{\Psi}_2$ is proportional to the product $L(\boldsymbol{\Psi}_2 \mid \{\mathbf{m}_i\}, \boldsymbol{\Lambda}_2, \boldsymbol{\mu}, \boldsymbol{\Theta}_2,$

$\{\boldsymbol{\delta}_{2,i}\})\pi(\text{vec}(\boldsymbol{\Psi}_2))$. Here $L(\cdot)$ is the conditional likelihood of observing $\{\mathbf{m}_i\}$ given the matrix $\boldsymbol{\Psi}_2$ and other parameters, and is proportional to

$$|\boldsymbol{\Theta}_2|^{-\frac{I}{2}} \exp\left\{-\tfrac{1}{2}\sum_{i=1}^{I}(\mathbf{m}_i - \boldsymbol{\mu} - \boldsymbol{\Lambda}_2\boldsymbol{\delta}_{2,i})'\boldsymbol{\Theta}_2^{-1}(\mathbf{m}_i - \boldsymbol{\mu} - \boldsymbol{\Lambda}_2\boldsymbol{\delta}_{2,i})\right\}. \tag{A11}$$

As direct methods for sampling from the full conditional are not available, we use the Metropolis Hit and Run algorithm. The acceptance probability can be constructed easily by substituting the appropriate prior and likelihood in the expression for the acceptance probability in (A2).

References

Albert, J., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association, 88*, 669–679.

Albert, J., & Chib, S. (1995). Bayesian residual analysis for binary response regression models. *Biometrika, 82*, 747–759.

Arminger, G., & Muthén B. (1998). A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis–Hastings algorithm. *Psychometrika, 63*, 271–300.

Bartholomew, D.J. (1980). Factor analysis for categorical data. *Journal of the Royal Statistical Society, Series B, 42*, 293–321.

Bartholomew, D.J. (1981). Posterior analysis of the factor model. *British Journal of Mathematical and Statistical Psychology, 34*, 93–99.

Bartholomew, D.J. (1984). Scaling binary data using a factor model. *Journal of the Royal Statistical Society, Series B, 46*, 120–123.

Bartholomew, D.J. (1987). *Latent variable models and factor analysis*, New York, NY: Oxford University Press.

Best, N.G., Cowles, M.K., & Vines, S.K. (1995). *CODA: Convergence diagnostics and output analysis software for Gibbs sampler output, Version 0.3.* (Tech. Rep.). Cambridge, UK: Biostatistics Unit-MRC.

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443–445.

Bock, R.D., & Gibbons, R.D., (1996). High-dimensional multivariate probit analysis. *Biometrics, 52*, 1183–1194.

Brooks, S.P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics, 7*, 434–455.

Brooks, S.P., & Roberts, G.O. (in press). Assessing convergence of Markov chain Monte Carlo algorithms. *Journal of Computational and Graphical Statistics.*

Chambers, R.G. (1982). Correlation coefficients from 2×2 tables and from biserial data. *British Journal of Mathematical and Statistical Psychology, 35*, 216–227.

Chen, Ming-Hui, & Dey, D.K. (1998). Bayesian modeling of correlated binary responses via scale mixture of multivariate normal link functions. *Sankhya, Series A, 60*, 322–343.

Chen, Ming-Hui, & Schmeiser, B.W. (1993). Performance of the Gibbs, Hit-and-Run, and Metropolis Samplers. *Journal of Computational and Graphical Statistics, 2*, 251–272.

Chib, S., & Greenberg, E. (1995). Understanding the Metropolis–Hastings Algorithm. *American Statistician, 49*, 327–35.

Chib, S., & Greenberg, E. (1998). Analysis of Multivariate Probit Models. *Biometrika, 85*(2), 347–361.

Christofferson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika, 40*, 5–32.

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.

Crosswhite, F.J., Dossey, J.A., Swafford, J.O., McKnight, C.C., & Cooney, T.J. (1985). *Second International Mathematics Study: Summary report for the United States.* Champaign, IL: Stipes.

Chen, M.H., & Dey, Dipak K. (1998). Bayesian analysis of correlated binary data models. *Sankhya, Series A, 60*, 322–343.

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

Gelfand, A.E., & Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association, 85*, 972–985.

Gelman, A., Carlin, J.B., Stern, H. S., & Rubin, D. R. (1996). Posterior predictive assessment of model fitness (with discussion). *Statistica Sinica, 6*, 733-807.

Gelman, A., & Rubin, D.R. (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 of Pattern Analysis and Machine Intelligence, 6*, 721–741.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments (with discussion). In J. M. Bernardo et al. (Eds.), *Bayesian Statistics 4* (pp. 156–163). Oxford: Oxford University Press.

Geyer, C.J. (1992). Practical Markov chain Monte Carlo (with discussion). *Statistical Science, 7*, 473–511.

Goldstein, H., & McDonald, R.P. (1988). A general model for the analysis of multilevel data. *Psychometrika*, 53, 455–467.

Kass, R.E., & Raftery, A.E. (1995). Bayes factors. *Journal of American Statistical Association, 90*, 773–795.

Lee, S.-Y (1981). A Bayesian approach to confirmatory factor analysis. *Psychometrika, 46*, 153–160.
Longford, N.T., & Muthén, B. (1992). Factor analysis for clustered observations. *Psychometrika, 57*, 581–597.
Mardia, K.V. (1970). *Families of bivariate distributions,* London: Griffin.
Martin, J.K. & McDonald, R.P. (1975). Bayesian estimation in unrestricted factor analysis; a treatment for Heywood cases. *Psychometrika, 40*, 505–517.
McDonald, R.P., & Goldstein, H. (1989). Balanced versus unbalanced designs for linear structural relations in two-level data. *British Journal of Mathematical and Statistical Psychology, 42*, 214–232.
Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika, 43*, 551–560.
Muthén, B. (1979). A structural probit model with latent variables. *Journal of the American Statistical Association, 74*, 807–811.
Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika , 49*, 115–132.
Muthén, B. (1987). *LISCOMP: Analysis of linear structural equations with a comprehensive measurement model* (User's Guide). Mooresville, IN: Scientific Software.
Muthén, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54*, 557–585.
Muthén, B. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research 22*, 376–398.
Muthén, B., & Christofferson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika, 46*, 407–419.
Muthén, B., & Satorra, A. (1989). Multilevel aspects of varying parameters in structural models. In R. D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 87–99). New York, NY: Academic Press.
Sahu, S.K. (1998). Bayesian estimation and model choice in item response models (Tech. Rep.). Cardiff, Wales, UK: Cardiff University, School of Mathematics.
Shi, J., & Lee, S.-Y. (1997). A Bayesian estimation of factor score in confirmatory factor model with polytomous, censored or truncated data. *Psychometrika, 62*, 29–50.
Tanner, M.A., & Wong, W.H.(1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of American Statistical Association, 82*, 528–550.
Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion), *Annals of Statistics, 22*, 1701–1762.

Manuscript received 3 MAR 1998
Final version received 12 OCT 1999