Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2009, Article ID 537139, 18 pages
doi:10.1155/2009/537139

Review Article

The Rise of Markov Chain Monte Carlo Estimation for Psychometric Modeling

Roy Levy

Division of Advanced Studies in Learning, Technology and Psychology in Education, Arizona State University, PO Box 870611, Tempe, AZ 85287-0611, USA

Correspondence should be addressed to Roy Levy, roy.levy@asu.edu

Received 7 August 2009; Accepted 26 November 2009

Recommended by Peter van der Heijden

Markov chain Monte Carlo (MCMC) estimation strategies represent a powerful approach to estimation in psychometric models. Popular MCMC samplers and their alignment with Bayesian approaches to modeling are discussed. Key historical and current developments of MCMC are surveyed, emphasizing how MCMC allows the researcher to overcome the limitations of other estimation paradigms, facilitates the estimation of models that might otherwise be intractable, and frees the researcher from certain possible misconceptions about the models.

1. Introduction

The last decade has seen an explosion in the use of Markov chain Monte Carlo (MCMC) techniques in fitting statistical psychometric models. In this time, MCMC has been put to advantageous use in estimating existing models and, more importantly, supporting the development of new models that are otherwise computationally intractable. This paper surveys the use of MCMC in modern psychometric models, namely, models that employ (a) probabilistic reasoning in the form of statistical models to facilitate inference from observations of behaviors made by subjects to more broadly conceived statements about the subjects and/or the domain and (b) latent variables to model the presence of measurement error. Additional modeling archetypes, including hierarchical and mixture modeling, are noted where they intersect or overlay with the psychometric modeling paradigm of interest.

Psychometric models are typically organized in terms of assumptions about the latent and observable variables. Factor analysis (FA; Bollen [1]; Gorsuch [2]) specifies continuous latent and observable variables and frequently additionally assumes the latter to be normally distributed. Structural equation modeling (SEM; Bollen [1]) may be conceptualized as extending the factor analytic tradition to include regression-like structures that relate latent variables to one another. Like FA, item response theory (IRT; Lord

[3]) assumes the latent variables to be continuous but assumes that the observables are discrete and, when polytomous, usually ordered. Latent class analysis (LCA; Lazarsfeld and Henry [4]) and related models assume that both the latent and observable variables are discrete.

This list is far from exhaustive and in later sections we will discuss these and other psychometric models, some of which can be viewed as extensions or combinations of those already mentioned. It is however important to recognize that these latent variable modeling traditions evolved somewhat independently from one another, yielding a current state in which a repository of fairly mature models possess only at best partially overlapping foci, literatures, notational schemes, and—of interest in this work—paradigmatic estimation routines and strategies. To illustrate, FA and SEM have historically been employed to model relationships among constructs and typically do not involve inferences at the subject level. Estimation typically involves maximum likelihood (ML) or least squares (LS) using first- and second-order moments from sample data, with an emphasis on the estimation of structural parameters, that is, parameters for the conditional distributions of observed scores given latent variables (e.g., factor loadings) but not on the values of the latent variables (factors) for individual subjects (Bollen, [1]). In contrast, IRT models are commonly employed to scale test items and subjects. Estimation is usually conducted using individual level data or frequencies of response patterns and assumptions regarding the distribution of the latent variables for subjects. As in FA and SEM, the estimation of structural parameters (here interpreted as item parameters) is important. However, in IRT the estimation of subjects' values for the latent variable(s) is important to guide desired inferences about subjects (Lord [3]). Disparate estimation approaches organized around differences in input data (moments versus raw data) and differences in the targets of the inference evolved in IRT and in FA and SEM, with the unfortunate consequence of obscuring fundamental similarities among the models and, as discussed below, hampering the development of each.

This paper is organized as follows. A brief description of the Bayesian approach to psychometric modeling is advanced, followed by an overview of the most popular MCMC samplers for psychometric models, where the emphasis is placed on how the elements of the latter align with the features and challenges of estimating posterior distributions in the former. Next, an overview of the key historical developments and current work on MCMC for psychometric modeling is presented, emphasizing how MCMC overcomes the limitations of other estimation paradigms, facilitates the estimation of models that would be otherwise intractable, and frees the researcher from possible misconceptions about the models. A discussion concludes the paper.

2. Bayesian Psychometric Modeling

Because MCMC procedures yield empirical approximations to probability distributions, it fits naturally with the Bayesian approach to statistical analysis in which unknown parameters are treated as random and represented with probability distributions (Gelman et al. [5]), though we note that MCMC estimation has been employed in frequentist applications as well (Song and Lee [6]). As will be highlighted below, key features of the most flexible of MCMC algorithms may be viewed as explicitly resolving the most difficult challenges in estimating Bayesian models. Quite aside from the issue of estimation, a Bayesian approach in which models are formulated hierarchically, prior information can be easily incorporated, and uncertainty in unknown parameters is propagated offers advantages regardless of the estimation routine (see, e.g., Lindley and Smith [7]; Mislevy [8], for illustrative applications not involving MCMC).

To formulate a Bayesian psychometric model, let $\boldsymbol{\theta}_i$ be a possibly vector-valued latent variable and let $\mathbf{X}_i = (X_{i1}, \ldots, X_{iJ})$ be a vector of J observable variables for subject i, $i = 1, \ldots, N$. Let $\boldsymbol{\theta}$ and $\mathbf{X}$ denote the full collections of latent and observable variables, respectively. Psychometric models are typically viewed in terms of their assumptions regarding $\boldsymbol{\theta}$, $\mathbf{X}$, and the specification of the probabilistic dependence of the latter on the former via a conditional distribution $P(\mathbf{X} \mid \boldsymbol{\theta}, \boldsymbol{\omega})$ where $\boldsymbol{\omega}$ are parameters that govern the conditional distribution of the data (e.g., factor loadings in FA/SEM, item parameters in IRT, class-specific conditional probabilities in LCA). Assumptions of subject and local independence imply that the distribution for any observable depends only on the latent variable(s) for that subject and the conditional distribution parameters specific to that observable. This allows for the factorization of the joint conditional distribution of $\mathbf{X}$ as

$$P(\mathbf{X} \mid \boldsymbol{\theta}, \boldsymbol{\omega}) = \prod_{i=1}^{N} \prod_{j=1}^{J} P\left(X_{ij} \mid \boldsymbol{\theta}_i, \boldsymbol{\omega}_j\right), \tag{2.1}$$

where $\boldsymbol{\omega}_j$ are the parameters of $\boldsymbol{\omega}$ that concern observable j.

To conduct Bayesian inference, we specify prior distributions for the unknown parameters. An assumption of exchangeability (Lindley and Smith [7]; Lindley and Novick [9]) implies that a common prior distribution may be used for all subjects' latent variables:

$$\boldsymbol{\theta}_i \sim P(\boldsymbol{\theta}_i \mid \boldsymbol{\eta}), \tag{2.2}$$

where $\boldsymbol{\eta}$ are higher level parameters that govern the distribution of $\boldsymbol{\theta}_i$. Depending on the application, $\boldsymbol{\eta}$ may serve the varied purposes of (a) resolving indeterminacies in the location, scale, and orientation of latent axes, (b) defining features of the population, or (c) specifying prior expectations. $\boldsymbol{\eta}$ may include free parameters that require estimation (e.g., factor covariances and structural coefficients in SEM); in these cases, these elements of $\boldsymbol{\eta}$ would be assigned a prior distribution $P(\boldsymbol{\eta})$.

Turning to $\boldsymbol{\omega}$, an exchangeability assumption with respect to observables implies that a common prior distribution may be used for all the conditional probability parameters for a given observable

$$\boldsymbol{\omega}_j \sim P(\boldsymbol{\omega}_j \mid \boldsymbol{\lambda}), \tag{2.3}$$

where $\boldsymbol{\lambda}$ are higher level parameters that govern the distribution of $\boldsymbol{\omega}_j$. In models that include multiple elements in $\boldsymbol{\omega}_j$, it is common to specify independent prior distributions for each element. For example, in three-parameter IRT models (Lord [3]), $\boldsymbol{\omega}_j = (b_j, a_j, c_j)$ and $P(\boldsymbol{\omega}_j \mid \boldsymbol{\lambda}) = P(b_j \mid \boldsymbol{\lambda}_b)P(a_j \mid \boldsymbol{\lambda}_a)P(c_j \mid \boldsymbol{\lambda}_c)$ where $\boldsymbol{\lambda}_b$, $\boldsymbol{\lambda}_a$, and $\boldsymbol{\lambda}_c$ are components of $\boldsymbol{\lambda}$ that govern the distributions of the b, a, and c parameters, respectively. When in need of estimation, elements of $\boldsymbol{\lambda}$ are assigned a prior distribution $P(\boldsymbol{\lambda})$.

Following the conditional independence assumptions inherent in the above treatment, the joint probability distribution for all the entities in the model can be expressed as

$$\begin{aligned} P(\mathbf{X},\boldsymbol{\theta},\boldsymbol{\omega},\boldsymbol{\eta},\boldsymbol{\lambda}) &= P(\mathbf{X} \mid \boldsymbol{\theta},\boldsymbol{\omega})P(\boldsymbol{\theta} \mid \boldsymbol{\eta})P(\boldsymbol{\omega} \mid \boldsymbol{\lambda})P(\boldsymbol{\eta})P(\boldsymbol{\lambda}) \\ &= \prod_{i=1}^{N}\prod_{j=1}^{J} P\left(X_{ij} \mid \boldsymbol{\theta}_i,\boldsymbol{\omega}_j\right)P\left(\boldsymbol{\theta}_i \mid \boldsymbol{\eta}\right)P\left(\boldsymbol{\omega}_j \mid \boldsymbol{\lambda}\right)P(\boldsymbol{\eta})P(\boldsymbol{\lambda}). \end{aligned} \tag{2.4}$$

Once values for $\mathbf{X}$ are observed, the posterior distribution for the unknown parameters is obtained as

$$\begin{aligned} P(\boldsymbol{\theta},\boldsymbol{\omega},\boldsymbol{\eta},\boldsymbol{\lambda} \mid \mathbf{X}) &= \frac{P(\mathbf{X} \mid \boldsymbol{\theta},\boldsymbol{\omega})P(\boldsymbol{\theta} \mid \boldsymbol{\eta})P(\boldsymbol{\omega} \mid \boldsymbol{\lambda})P(\boldsymbol{\eta})P(\boldsymbol{\lambda})}{P(\mathbf{X})} \\ &= \frac{P(\mathbf{X} \mid \boldsymbol{\theta},\boldsymbol{\omega})P(\boldsymbol{\theta} \mid \boldsymbol{\eta})P(\boldsymbol{\omega} \mid \boldsymbol{\lambda})P(\boldsymbol{\eta})P(\boldsymbol{\lambda})}{\int_{\boldsymbol{\theta}}\int_{\boldsymbol{\omega}}\int_{\boldsymbol{\eta}}\int_{\boldsymbol{\lambda}} P(\mathbf{X} \mid \boldsymbol{\theta},\boldsymbol{\omega})P(\boldsymbol{\theta} \mid \boldsymbol{\eta})P(\boldsymbol{\omega} \mid \boldsymbol{\lambda})P(\boldsymbol{\eta})P(\boldsymbol{\lambda})d\boldsymbol{\theta}d\boldsymbol{\omega}\mathrm{d}\boldsymbol{\eta}d\boldsymbol{\lambda}} \\ &\propto P(\mathbf{X} \mid \boldsymbol{\theta},\boldsymbol{\omega})P(\boldsymbol{\theta} \mid \boldsymbol{\eta})P(\boldsymbol{\omega} \mid \boldsymbol{\lambda})P(\boldsymbol{\eta})P(\boldsymbol{\lambda}). \end{aligned} \tag{2.5}$$

The current treatment resulting in the expression in (2.5) represents a basic psychometric model. Extensions are possible and indeed are frequently warranted. For example, the exchangeability assumptions for $\boldsymbol{\theta}_i$ (similarly, $\boldsymbol{\omega}_j$) may not be warranted if subjects (observables) are hierarchically structured. In such cases, an assumption of conditional exchangeability implies the use of group-specific prior distributions, possibly involving covariates (De Boeck and Wilson [10]). In passing, we note that MCMC handles such hierarchical or conditional structures in a straightforward manner; examples of MCMC applications for these and other extensions will be discussed below. For the current purposes, the development of (2.5) is sufficient to motivate the discussion of the development and advantages of MCMC.

3. Markov Chain Monte Carlo Estimation

Model estimation comes to estimating the posterior distribution. Analytical solutions, though ideal, are often impractical or impossible due to the necessity to evaluate the high-dimensional integrals to obtain the marginal distribution in the denominator in (2.5). When the posterior distribution is of known form, an empirical approximation may be obtained by simulating values using straightforward Monte Carlo procedures (Lee [11]). However, drawing independent samples is often computationally intractable, as posterior distributions in psychometric models are most often not of known form. MCMC (Brooks [12], Gelfand and Smith[13]; Gilks et al. [14]; Smith and Roberts [15]; Spiegelhalter et al. [16]; Tierney [17]) estimation consists of drawing possibly dependent samples from a distribution of interest and as such provides an appropriate framework for computation in Bayesian analyses (Gelman et al. [5]). Broadly speaking, we construct a Markov chain that has the posterior distribution as its stationary distribution; that is, MCMC estimation consists of drawing from a series of distributions that is in the limit equal to drawing from the stationary (posterior) distribution (Gilks et al. [14]).

To construct a Markov chain, initial values for all parameters must be specified. Subsequent values for the parameters are repeatedly drawn creating a sequence that constitutes the chain. Given certain general conditions hold (e.g., Roberts [18]), a properly constructed chain is guaranteed to converge to its stationary or target distribution.

In the following sections, several of the most popular MCMC routines are described. Emphasis is placed on the connection between the features of the MCMC routines that align with features of Bayesian approaches to psychometric modeling.

3.1. Gibbs Sampling

Let $\mathbf{\Omega} = (\Omega_1, \Omega_2, \ldots, \Omega_R)$ denote the R parameters in the model and let $\mathbf{X}$ denote the observed data. Let $P(\Omega_r \mid \mathbf{X}, \mathbf{\Omega}_{(r)})$ denote the *full conditional distribution* of the rth model parameter, the conditional distribution of the parameter given the data ($\mathbf{X}$) and all other model parameters $\mathbf{\Omega}_{(r)}$. It can be shown that a joint distribution may be defined by the complete set of such full conditional distributions (Besag [19]; Gelfand and Smith [13]). Thus in a Bayesian analysis, the joint posterior distribution of model parameters may be defined as the complete set of full conditional posterior distributions. That is, the joint posterior $P(\mathbf{\Omega} \mid \mathbf{X})$ may be defined by $P(\Omega_1 \mid \mathbf{X}, \mathbf{\Omega}_{(1)}), P(\Omega_2 \mid \mathbf{X}, \mathbf{\Omega}_{(2)}), \ldots, P(\Omega_R \mid \mathbf{X}, \mathbf{\Omega}_{(R)})$. Sampling from the joint posterior then comes to sampling from these full conditional distributions.

Let ${\Omega_r}^t$ denote the value of model parameter r at iteration t. Gibbs sampling (Gelfand and Smith [13]; S. Geman and D. Geman [20]; see also Brooks [12]; Casella and George [21] ; Gilks et al. [14]) consists of proceeding to the following steps.

(1) Initialize the parameters by assigning values for $\Omega_1^t, \Omega_2^t, \ldots, \Omega_R^t$ for $t = 0$.

(2) For $r = 1, \ldots, R$, draw values for parameter Ω_r from its full conditional distribution given the current values of all other model parameters and the observed data. That is, for each parameter Ω_r, we obtain the $t+1$st iteration value of the chain by drawing from $P(\Omega_r \mid \mathbf{X}, \Omega_1^{t+1}, \ldots, \Omega_{r-1}^{t+1}, \Omega_{r+1}^t, \ldots, \Omega_R^t)$. One cycle is given by sequentially drawing values from

$$\begin{gathered}
\Omega_1^{t+1} \sim P(\Omega_1 \mid \mathbf{X}, \Omega_2^t, \ldots, \Omega_R^t) \\
\Omega_2^{t+1} \sim P\Big(\Omega_2 \mid \mathbf{X}, \Omega_1^{t+1}, \Omega_3^t, \ldots, \Omega_R^t\Big) \\
\vdots \\
\Omega_R^{t+1} \sim P\Big(\Omega_R \mid \mathbf{X}, \Omega_1^{t+1}, \Omega_3^t, \ldots, \Omega_{R-1}^{t+1}\Big).
\end{gathered} \tag{3.1}$$

(3) Repeat step 2 for some large number T iterations.

The conditional independence assumptions greatly reduce the set of parameters that need to be conditioned on in each of distributions in step 2. For example, in drawing a value for a subject's value of $\boldsymbol{\theta}$, respondent independence implies that the values of $\boldsymbol{\theta}$ for the remaining subjects need not be considered. Note also in step 2 that each draw for iteration $t+1$ is subsequently used in the full conditionals for the remaining parameters.

3.2. Metropolis-Hastings Sampling

In complex models, it may be the case that while full conditional distributions may be constructed, they are too complex to sample from. More complex sampling schemes, such as the Metropolis-Hastings and Metropolis samplers, described in this and the following section, are required.

To simplify notation, let $\pi(\mathbf{\Omega}) = P(\mathbf{\Omega} \mid \mathbf{X})$ denote the target distribution (i.e., the posterior distribution of interest). Metropolis-Hastings sampling (Hastings [22]; see also Brooks [12]; Chib and Greenberg [23]; Gilks et al. [14]) consists of conducting the following steps.

(1) Initialize the parameters by assigning a value for $\mathbf{\Omega}^t$ for $t = 0$.

(2) Draw a *candidate value* $\mathbf{y} \sim q(\mathbf{y} \mid \mathbf{\Omega}^t)$ from a *proposal* distribution q.

(3) Accept $\mathbf{y}$ as the $t+1$st iteration for $\mathbf{\Omega}$ with probability $\alpha(\mathbf{y} \mid \mathbf{\Omega}^t) = \min[1, \pi(\mathbf{y})q(\mathbf{\Omega}^t \mid \mathbf{y})/\pi(\mathbf{\Omega}^t)q(\mathbf{y} \mid \mathbf{\Omega}^t)]$. Retain the current value of $\mathbf{\Omega}^t$ for $\mathbf{\Omega}^{t+1}$ with probability $1-\alpha(\mathbf{y} \mid \mathbf{\Omega}^t)$.

(4) Repeat steps 2 and 3 for some large number T iterations.

The *acceptance probability* $\alpha(\mathbf{y} \mid \mathbf{\Omega}^t)$ involves evaluating the posterior distribution π and the proposal distribution q at both the current and candidate values. Note that in the formulation here the proposal distribution q may be conditional on the current value of the chain, which constitutes a random-walk sampler (Brooks [12]). More generally, q may be any distribution that is defined over the support of the stationary distribution π. As such, the Metropolis-Hastings algorithm is an extremely flexible approach to estimating posterior distributions.

3.3. Metropolis Sampling

In Metropolis sampling (Metropolis et al. [24]; see also Brooks [12]; Gilks et al. [14]), q is chosen so that it is *symmetric with respect to its arguments*, $q(\mathbf{y} \mid \mathbf{\Omega}^t) = q(\mathbf{\Omega}^t \mid \mathbf{y})$. The acceptance probability then simplifies to $\alpha(\mathbf{y} \mid \mathbf{\Omega}^t) = \min[1, \pi(\mathbf{y})/\pi(\mathbf{\Omega}^t)]$. A popular choice for q is the normal distribution centered at the current value of the chain.

It is easily seen that the Metropolis sampler is a special case of the Metropolis-Hastings sampler. It is somewhat less obvious that the Gibbs sampler may be viewed as a special case of the Metropolis sampler, namely, where the proposal distribution for each parameter is the full conditional distribution, which implies that the acceptance probability α will equal 1.

Recall that in Bayesian analyses of psychometric models the posterior distribution is generally only known up until a constant of proportionality (see (2.5)). Further, recall that we construct the chain to have the posterior distribution of interest as the stationary distribution. Inspection of the Metropolis(-Hastings) sampler(s) reveals that the stationary distribution π (i.e., the posterior distribution) appears in both the numerator and denominator of the acceptance probability α and therefore only needs to be known up to a constant of proportionality. MCMC alleviates the need to conduct high-dimensional integration over the parameter space to estimate the posterior distribution. This is the key feature of MCMC estimation that permits the estimation of complex Bayesian models.

3.4. Metropolis(-Hastings)-Within-Gibbs

A Metropolis(-Hastings)-within-Gibbs sampler, also termed single-component-Metropolis (-Hastings), combines the component decomposition approach of the Gibbs sampler with the flexibility of Metropolis(-Hastings). As noted above, Gibbs sampling involves sampling from the full conditional distributions for each parameter separately. When these full conditionals are not of known form, a Metropolis(-Hastings) step may be taken where, for each parameter, a candidate value is drawn from a proposal distribution q and accepted with probability α as the next value in the chain for that parameter.

4. Psychometric Modeling Using MCMC

This section reviews key developments in the growing literature on psychometric modeling using MCMC. In tracing the foundational developments and current applications, the emphasis is placed on models and modeling scenarios where the power of MCMC is leveraged to facilitate estimation that would prove difficult of not intractable for traditional procedures, highlighting the flexibility of MCMC and the resulting freedom it provides.

4.1. Continuous Latent and Observable Variables

FA and SEM models with linear equations relating the latent and observed variables and (conditional) normality assumptions may be easily handled by traditional ML or LS estimation routines (Bollen [1]). However, in introducing Gibbs sampling schemes for SEM, Scheines et al. [25] noted that Gibbs sampling holds a number of advantages in that (a) it does not rely on asymptotic arguments for estimation or model checking and therefore may be better suited for small samples (Ansari and Jedidi [26]; Lee and Song [27]), (b) inequality constraints may be easily imposed, (c) information about multimodality—undetectable by standard ML estimation—may be seen in marginal posterior densities, and (d) information for underidentified parameters may be supplied via informative priors.

The great advantage of MCMC for SEM lies in its power to estimate nonstandard models that pose considerable challenges for ML and LS estimation (Lee [11]). Examples of such applications include heterogeneous and multilevel factor analysis models (Ansari et al. [28]), complex growth curve models (Zhang et al. [29]), latent mixture SEM (Lee and Song [30]; Zhu and Lee [31]), and models with covariates (Lee et al. [32]), nonignorable missingness (Lee and Tang [33]), or interaction, quadratic, and similarly nonlinear relationships among latent variables including nonlinear longitudinal effects (Arminger and Muthén [34]; Lee et al. [32]; Lee and Tang [33]; Song et al. [35]).

The implication is that, though traditional estimation routines that evolved with the standard FA and SEM paradigm may be suitable for simple models, extending the standard models to more complex situations may necessitate the use of more flexible MCMC procedures. Moreover, contrary to a common belief, the computation necessary to implement MCMC estimation in such complex models is generally less intense than that necessary to conduct ML estimation (Ansari and Jedidi [26]; Ansari et al. [28]; Zhang et al. [29]).

4.2. Continuous Latent Variables and Discrete Observable Variables

In this section, we survey applications of MCMC to models in which a set of discrete, possibly ordinal observables are structured as indicators of continuous latent variables from both FA and IRT perspectives, highlighting aspects in which existing estimation traditions limit our modeling potential.

The FA tradition models discrete data using continuous latent variables by considering the observables to be discretized versions of latent, normally distributed data termed latent response variables. Traditional FA estimation methods have relied on calculating and factoring polychoric correlations, which involves the integration over the distribution of the latent response variables. This approach suffers in that the FA routines were explicitly developed for continuous rather than discrete data. Wirth and Edwards [36] concluded that traditional FA methods can fail to capture the true fit of the model, even with corrections to estimates and standard errors for discrete data. This illustrates the limitations that analysts encounter by remaining within an estimation paradigm when trying to fit models beyond the scope of those originally intended for the estimation routine.

The dominant estimation paradigm in IRT involves marginal maximum likelihood (MML; Bock and Aitkin [37]; see also Baker and Kim [38]), in which the marginal distribution of the data as a function of item parameters is produced by numerically integrating over an assumed distribution of the latent continuous variables. Taking this function as a marginal likelihood for the item parameters, estimates for item parameters are obtained by maximizing this function, possibly augmented by prior distributions (Mislevy [8]). In assessment scenarios, estimation of subject parameters is of interest to facilitate subject-level inferences. This is conducted by treating the just-estimated item parameters as known to produce a likelihood function for the subject parameters, which is either (a) maximized or (b) maximized or averaged over after being augmented by a prior distribution (e.g., Bock and Mislevy [39]). This divide-and-conquer strategy suffers in that uncertainty in the estimation of item parameters in the first stage is ignored when estimating subjects' parameters. What is needed—and what MCMC provides—is an estimation framework flexible enough to handle a variety of assumptions about the distributional features of the observable variables, latent variables, and the data-generating process, not to mention the all-too-real potential for missingness or sparseness, all the while properly propagating uncertainty throughout.

A foundation for MCMC for IRT, and psychometric modeling more generally, was given by Albert [40], whose seminal work showed how posterior distributions for item and subject parameters in normal-ogive IRT models could be estimated via Gibbs sampling using data augmentation strategies (Tanner and Wong [41]). The algorithm was extended to handle polytomous data by Albert and Chib [42]; a similar Gibbs sampling approach was described by Sahu [43] that allows for guessing as may be applicable in assessment contexts.

The turning point in the application of MCMC for psychometric modeling came with the work of Patz and Junker [44], who offered a Metropolis-Hastings-within-Gibbs sampling approach for the most common logistic IRT models. The flexibility of this approach has produced an explosion in the use of MCMC for IRT-based models, including those for polytomous-ordered data (Patz and Junker [45]), nominal data (Wollack et al. [46]), missingness (Patz and Junker [45]), rater effects (Patz and Junker [45]), testlets (Bradlow et al. [47]), multilevel models (Fox and Glas [48]), and hierarchical models for mastery classification (Janssen et al. [49]).

To highlight an arena where the intersection of different modeling paradigms and their associated traditional estimation routines poses unnecessary limits, consider multidimensional models for discrete observables. The equivalence between IRT and FA versions of the models has long been recognized (Takane and de Leeuw [50]). However, as noted by Wirth and Edwards [36], common misconceptions associated with each perspective can be tied to the historical estimation traditions within each paradigm. In the FA tradition, the calculation of polychoric correlations involves the integration over the distribution of the latent response variables. This integration becomes increasingly difficult as the number of observables increases (as each observable has its own latent response variable), and the applicability of ML and weighted LS routines requiring large sample sizes relative to the number of observables becomes suspect. As a consequence, the FA perspective prefers (relatively) few observables in the model but has no concern for the number of latent variables. In contrast, traditional MML estimation approaches in IRT perform an integration over the latent variables, which becomes increasingly difficult as the number of latent variables increases. As a consequence, the IRT perspective prefers (relatively) few latent variables in the model but is ambivalent toward the number of observables. Despite the long-standing recognition of the equivalence of these perspectives with respect to the *model*, the adoption of either one or the other tradition-specific *estimation* paradigms restricts the scope of the model's application.

MCMC may be seen as a unifying framework for estimation that frees the analyst from these restrictive—and conflicting—misconceptions. Examples of the use of MCMC in the multidimensional modeling from both FA and IRT perspectives include the consideration of dichotomous data (Béguin and Glas [51]; Bolt and Lall [52]; Jackman [53]; Lee and Song [30]), polytomous data (Yao and Boughton [54]), and combinations of continuous, dichotomous, and polytomous data (Lee and Zhu [55]; Shi and Lee [56]), as well as models for multiple groups (Song and Lee [57]), missing data (Song and Lee [58]), nonlinear relationships among latent variables (Lee and Zhu [55]), and multilevel structures (Ansari and Jedidi [26]; Fox and Glas [48]).

4.3. Discrete Latent Variables and Discrete Observable Variables

Similar to the case of FA of continuous data, ML estimation can typically handle traditional, unrestricted LCA models that model discrete observables as dependent on discrete latent variables. And here again, MCMC may still be advantageous for such models in handling missingness, large data sets with outliers, and constructing credibility intervals for inference when an assumption of multivariate normality (of ML estimates or posterior distributions) is unwarranted (Hoijtink [59]; Hoijtink and Notenboom [60]). Similar to the case of IRT, traditional estimation in LCA proceeds with a divide-and-conquer approach in which conditional probabilities are estimated in one stage and then treated as known in a second stage to estimate subject parameters. As noted above, MCMC simultaneously estimates all parameters all the while properly accounting for the uncertainty in the estimation.

Turning to more complex models, MCMC has proven useful in estimating models with covariates (Chung et al. [61]) and with ordinal and inequality constraints (van Onna [62]). In assessment scenarios, diagnostic classification models (Rupp and Templin [63]) typically model discrete observables (i.e., scored item responses) as dependent on different combinations of the latent, typically binary, attributes characterizing mastery of componential skills necessary to complete the various tasks. The models frequently involve conjunctive or disjunctive relationships to model the probabilistic nature of student responses. These models pose estimation difficulties for traditional routines but can be handled by MCMC (de la Torre and Douglas [64]; Hartz [65]; Henson et al. [66]; Templin and Henson [67]).

These models may be also be cast as Bayesian networks, which allow for the estimation of a wide variety of complex effects via MCMC. Examples include compensatory, conjunctive, disjunctive, and inhibitor relationships for dichotomous and polytomous data assuming for dichotomous or ordered latent student skills or attributes (Almond et al. [68]; Levy and Mislevy [69]). The recent growth in interest in these and other models that attempt to more accurately depict the structures and processes of human reasoning (see, e.g., Bolt and Lall [52], on the use of MCMC to fit conjunctive multidimensional IRT models) illustrates how the flexibility of MCMC opens the door for the application of complex statistical models that are more closely aligned with substantive theories regarding the domain and the data-generating process.

4.4. Combinations of Models

The discussion has so far been couched in terms of traditional divisions between models, highlighting applications that pose difficulties for estimation routines typically employed. An advanced approach to model construction takes a modular approach in which the statistical model is constructed in a piecewise manner, interweaving and overlaying features from the traditional paradigms as necessary (Rupp [70]). Simple examples include the models that bridge the FA and IRT divided by modeling discrete and continuous observables simultaneously (Lee and Zhu [55]; Shi and Lee [56]). More nuanced examples embed IRT and FA models in latent classes to construct latent mixtures of IRT or FA models (Bolt et al. [71]; Cohen and Bolt [72]; Lee and Song [30]; Zhu and Lee [31]).

The machinery of MCMC can be brought to bear in addressing recurring complications inherent in psychometric applications. MCMC naturally handles missing data (e.g., Chung et al. [61]; Lee and Tang [33]; Patz and Junker [45]; Song and Lee [58]) and offers a unified strategy for handling latent variables as missing data (Bollen [73]; Jackman [74]). Similarly, recent work has sought to simultaneously address the hierarchical structures of data as well as the presence of measurement error. Examples of the use of MCMC for multilevel psychometric models can be found in Ansari and Jedidi [26], Ansari et al. [28], Fox and Glas [48], and Mariano and Junker [75]. To date, traditional estimation strategies have not been established for these models.

To illustrate the need for a comprehensive model estimation paradigm that is sensitive to the various data structures—and the capability of MCMC to fill that need—consider the National Assessment of Educational Progress (NAEP), which is characterized by (a) inferences targeted at the level of (sub)populations (rather than individuals) that are hierarchically organized, (b) administration of dichotomously and polytomously scored items, (c) complex sampling designs for subjects and items, and (d) covariates at each level of the analysis.

Beginning with the piecewise traditional approach, multiple-imputation approaches (Beaton [76]) accounted for the sampling design of subjects with jackknife procedures and used IRT to combine information across distinct booklets of items, but they suffered in that the point estimates of the IRT item parameters and latent regression models on covariates were treated as known. Scott and Ip [77] demonstrated a Bayesian framework for the multidimensional IRT model NAEP employs but do not consider the complex sampling design. Longford [78] and Raudenbush et al. [79] detail population-based analyses for data sets with hierarchical structures but did not address the presence of measurement error.

In contrast, Johnson and Jenkins [80] (see also Johnson, [81]) provided a Bayesian approach—estimated with MCMC—to model the (sub)population distributions accounting for the clustered-sampling designs and the matrix sampled item presentation. On the basis of analyses of simulated and real data from operational NAEP, Johnson and Jenkins

[80] compared the results from their unified model to the NAEP analysis with its piecewise approximations and found that both approaches provided consistent estimates of subpopulation characteristics, but their unified model more appropriately captured the variability of those estimates. By treating IRT item parameters and population variances as known, the standard analyses systematically underestimated the posterior uncertainty. Moreover, MCMC estimation of their unified model provided more stable estimates of sampling variability than the standard procedures. In this case, the use of MCMC estimation supported an analytic sampling and psychometric model that simultaneously better captured significant features of the design and provided better calibrated inferences for (sub)population characteristics of interest.

5. Discussion

The Gibbs, Metropolis-Hastings, and Metropolis samplers are described in the context of psychometric models with latent variables to illustrate the flexibility and power of MCMC in estimating psychometric models under a Bayesian paradigm. It is emphasized that the partitioning of the parameter space of Gibbs samplers and the requirement that the stationary distribution need only be specified up to a constant of proportionality in Metropolis (-Hastings) aligns these MCMC routines with the key features of the characteristics and challenges posed by the desired posterior distribution in Bayesian psychometric modeling.

Many of the examples are given to highlight how MCMC can be leveraged to (a) estimate complex statistical psychometric models that cannot be practically estimated by conventional means and (b) overcome the limitations of other approaches in situations in which the traditions of modeling and estimation paradigms unnecessarily restrict the scope of the models. Other examples highlight how MCMC can be gainfully employed in settings where alternative estimation routines already exist, such as in the analysis of small samples, missing data, and possible underidentification, and where divide-and-conquer strategies systematically understate the uncertainty in estimation.

Despite these advances, it is far from clear that MCMC will become as prevalent as, let alone replace, traditional likelihood-based or least-squares estimation. For straightforward applications of paradigmatic factor analytic, structural equation, item response, and latent class models, traditional estimation is fairly routine, accurate, and accessible via widely-available software (e.g., Mislevy and Bock [82], L. K. Muthén and B. O. Muthén [83]) and may outperform MCMC (see, e.g., Baker [84] in the context of popular IRT models). For more complex models, advances in tools for conducting traditional estimation (e.g., Schilling and Bock [85]) and broad perspectives on modeling (e.g., B. O. Muthén [86], Rabe-Hesketch et al. [87]) have supported the development of estimation routines and software for complicated and nuanced models (e.g., L. K. Muthén and B. O. Muthén [83]; Rabe-Hesketch et al. [88]). Nevertheless, when researchers want to push the boundaries of even these complex modeling paradigms, they may benefit by turning to MCMC (Segawa et al. [89]).

It is readily acknowledged that MCMC is difficult, both computationally in terms of necessary resources and conceptually in terms of constructing the chains, making relevant choices, and understanding the results. As to the computing challenge, the availability of software for conducting MCMC is a burgeoning area. Programs are available for conducting MCMC for IRT, FA, SEM, and diagnostic classification models (Arbuckle [90]; Henson et al. [66]; Jackman [91], Martin et al. [92]; Sheng [93, 94]; Yao [95]). Furthermore, general use software such as the freely available WinBUGS (Spiegelhalter et al. [16]) and its variants offers the psychometric community important resources for fitting familiar and innovative models not yet packaged elsewhere. The publishing of code for such software (e.g., Bolt and Lall [52]; Congdon [96]; Gill [97]; Lee [11]; Song et al. [35]) represents a meaningful step in making these programs more accessible. Similarly, companion software for interfacing MCMC software with other general statistical analysis software (e.g., Sturtz et al. [98]) and software for analyzing the output of MCMC (Plummer et al. [99]; Smith [100]) constitute more tools for researchers to employ when conducting MCMC estimation.

The criticism that MCMC is conceptually difficult is somewhat ironic, given that—for complex statistical models that reflect substantively rich hypotheses—it may actually be *easier* to set up an MCMC estimation routine than it is to proceed through the necessary steps (e.g., solving for first- and possibly second-order derivatives) in ML and LS estimation routines (Ansari and Jedidi [26]; Ansari et al. [28]; Zhang et al. [29]). MCMC allows analysts to estimate models without the requiring high-dimensional calculus necessary to obtain (a) derivatives in frequentist approaches to estimation or (b) the marginal distribution in a Bayesian approach (i.e., in the denominator of (2.5)).

Nevertheless, there is no debating that a certain level of technical sophistication is required to properly conduct an MCMC analysis. To this end didactic treatments of MCMC generally (Brooks [12]; Casella and George [21]; Chib and Greenberg [23]; Gilks et al. [14]; Jackman [101]) and in the context of psychometric models (Kim and Bolt [102]) constitute a firm foundation for researchers and practitioners learning about MCMC. The increasing number of published applications of MCMC for simple and complex psychometric analyses, especially those in textbook form that draw on the aforementioned software packages (Baker and Kim [38]; Congdon [96]; Gelman et al. [5], Gill [97]; Lee [11]), will also prove invaluable.

Further assistance in this area is provided by research that focuses on specific aspects of MCMC estimation. For example, the research community is aided by dedicated treatments and research on the complex issues of convergence assessment (see Sinharay [103] for a review and examples in psychometric applications) and data-model fit assessment and model comparisons. These latter two areas illuminate another potential unifying benefit of MCMC. Data-model fit assessment frameworks using traditional estimation procedures are varied, often localized to assumptions of the features of the observables or the aspects of fit, and not easily generalized across models (Bollen [1]; Swaminathan et al. [104]. With regard to this last point, the usage of modification indices across modeling paradigms represents a notable exception (Glas and Falćon [105]; Sörbom, [106]). Data-model fit based on traditional estimation may be limited in that it may be difficult to derive sampling distributions (e.g., applications of many popular discrepancy measures in SEM frequently involve a comparison to values to debated cutoff values, Hu and Bentler [107]). When sampling distributions are advanced, they may not be well defined (see, e.g., Chen and Thissen [108], for examples in IRT) and when they are well defined, they are typically justified only asymptotically.

The computations necessary for Bayesian procedures, including posterior predictive model checking (Gelman et al. [109]) and the Deviance Information Criterion (Spiegelhalter et al. [110]), are easily conducted using MCMC, do not depend on asymptotic arguments, fully propagate uncertainty in estimation, and are widely applicable across different kinds of models (for examples from a wide variety of psychometric models, see, e.g., Ansari et al. [28]; Fu et al. [111]; Lee [11]; Levy et al. [112]; Sahu [43]; Scheines et al. [25]; Sheng and Wikle [113]; Sinharay [103, 114], Sinharay et al. [115]).

This is not to assert that these approaches to data-model fit are necessarily superior to traditional approaches or are firmly established without debate. For example, posterior predictive checks (Gelman et al. [109]) have been advanced as a powerful and flexible approach to data-model fit assessment. However, these methods have been critiqued on the grounds that they yield P-values are not uniformly distributed under null conditions (Bayarri and Berger [116]; Robins et al. [117]), yielding conservative tests. This critique has spurned research aimed at identifying (a) situations in which will not be the case (Bayarri and Berger [118]; Gelman [119]; Johnson [120]), (b) mechanisms to calibrate the p-values (Hjort et al. [121]), and (c) alternative methods in the Bayesian/MCMC tradition that not are not subject to this critique (Bayarri and Berger [116]). Quite apart from this line of research lies an active debate regarding whether the critique is indeed problematic for Bayeisan modeling (Gelman [119, 122]). It remains to be seen if these methods of data-model fit assessment outperform those used in traditional approaches to psychometric modeling. Such a finding would be necessary if they are to be utilized in a more widespread manner. Additionally, didactic treatments of this and other aspects of MCMC will prove important to their increased usage. It is hoped that the current work adds to the growing body of resources by conceptually arguing why the computational features make MCMC so advantageous and evidencing its existing successes and potential for the future.

With regard to this last point, a number of historically reoccurring features, assumptions, and beliefs about psychometric models (e.g., linear relationships, independence and normality of errors, few latent variables in IRT, few discrete observables in FA) have evolved in part from limitations on estimation routines. The flexibility of MCMC frees the analyst from the bonds associated with other estimation approaches and allows the construction of models based on substantive theory. Indeed, the lasting impact of Patz and Junker's [44, 45] work on the generality of Metropolis-Hastings-within-Gibbs was not only that MCMC could be employed to estimate existing models of varying complexity but also that MCMC was a general approach to estimation flexible enough to handle any psychometric model that could be constructed. The explosion of MCMC in psychometrics in the past decade serves as a testament to this new state of affairs. Under this new paradigm, names like FA, IRT, and LCA no longer need to reflect choices that must be made about models—and possibly-limiting associated estimation procedures—but rather modules of recurring relationships or structures of associations that can be adapted and assembled to suit the substantive problem at hand (Rupp [70]). From this model-building perspective, it is worth noting that augmenting familiar models with multilevel or finite mixture structures poses complications for tradition estimation routines but are rather trivial increments for MCMC routines (Lee [11]). It is no surprise that analogous recommendations for modular model construction are found in the Bayesian literature (Gelman et al. [5]; Pearl [123]). A Bayesian framework, and in particular the power and flexibility of MCMC estimation, supports the removal of historical boundaries that are likely to hinder the growth of substantively rich and methodological complex psychometric models.

Acknowledgments

The author wishes to express his gratitude toward Bob Mislevy and John Behrens for stimulating conversations on the utility of MCMC. Thanks also to Bob Mislevy, Dubravka Svetina, and an anonymous reviewer for their helpful comments regarding earlier versions of the manuscript.

References

[1] K. A. Bollen, *Structural Equations with Latent Variables*, Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, New York, NY, USA, 1989.

[2] R. L. Gorsuch, *Factor Analysis*, Lawrence Earlbaum Associates, Hillsdale, NJ, USA, 2nd edition, 1983.

[3] F. M. Lord, *Applications of Item Response Theory to Practical Testing Problems*, Lawrence Earlbaum Associates, Hillsdale, NJ, USA, 1980.

[4] P. F. Lazarsfeld and N. W. Henry, *Latent Structure Analysis*, Houghton Mifflin, Boston, Mass, USA, 1968.

[5] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, Texts in Statistical Science Series, Chapman & Hall, London, UK, 1995.

[6] X.-Y. Song and S.-Y. Lee, "Full maximum likelihood estimation of polychoric and polyserial correlations with missing data," *Multivariate Behavioral Research*, vol. 38, no. 1, pp. 57–79, 2003.

[7] D. V. Lindley and A. F. M. Smith, "Bayes estimates for the linear model," *Journal of the Royal Statistical Society. Series B*, vol. 34, pp. 1–41, 1972.

[8] R. J. Mislevy, "Bayes modal estimation in item response models," *Psychometrika*, vol. 51, no. 2, pp. 177–195, 1986.

[9] D. V. Lindley and M. R. Novick, "The role of exchangeability in inference," *The Annals of Statistics*, vol. 9, no. 1, pp. 45–58, 1981.

[10] P. De Boeck and M. Wilson, Eds., *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*, Statistics for Social Science and Public Policy, Springer, New York, NY, USA, 2004.

[11] S.-Y. Lee, *Structural Equation Modeling: A Bayesian Approach*, Wiley Series in Probability and Statistics, John Wiley & Sons, Chichester, UK, 2007.

[12] S. P. Brooks, "Markov chain Monte Carlo method and its application," *The Statistician*, vol. 47, no. 1, pp. 69–100, 1998.

[13] A. E. Gelfand and A. F. M. Smith, "Sampling-based approaches to calculating marginal densities," *Journal of the American Statistical Association*, vol. 85, no. 410, pp. 398–409, 1990.

[14] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, Eds., *Markov Chain Monte Carlo in Practice*, Interdisciplinary Statistics, Chapman & Hall, London, UK, 1996.

[15] A. F. M. Smith and G. O. Roberts, "Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods," *Journal of the Royal Statistical Society. Series B*, vol. 55, no. 1, pp. 3–23, 1993.

[16] D. J. Spiegelhalter, A. Thomas, N. G. Best, and D. Lunn, *WinBUGS User Manual: Version 1.4.3*, MRC Biostatistics Unit, Cambridge, UK, 2007.

[17] L. Tierney, "Markov chains for exploring posterior distributions," *The Annals of Statistics*, vol. 22, no. 4, pp. 1701–1762, 1994.

[18] G. O. Roberts, "Markov chain concepts related to sampling algorithms," in *Markov Chain Monte Carlo in Practice*, W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, Eds., Interdiscip. Statist., pp. 45–57, Chapman & Hall, London, UK, 1996.

[19] J. Besag, "Spatial interaction and the statistical analysis of lattice systems," *Journal of the Royal Statistical Society. Series B*, vol. 36, pp. 192–236, 1974.

[20] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, no. 6, pp. 721–741, 1984.

[21] G. Casella and E. I. George, "Explaining the Gibbs sampler," *The American Statistician*, vol. 46, no. 3, pp. 167–174, 1992.

[22] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970.

[23] S. Chib and E. Greenberg, "Understanding the Metropolis-Hastings algorithm," *The American Statistician*, vol. 49, no. 4, pp. 327–335, 1995.

[24] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *The Journal of Chemical Physics*, vol. 21, no. 6, pp. 1087–1092, 1953.

[25] R. Scheines, H. Hoijtink, and A. Boomsma, "Bayesian estimation and testing of structural equation models," *Psychometrika*, vol. 64, no. 1, pp. 37–52, 1999.

[26] A. Ansari and K. Jedidi, "Bayesian factor analysis for multilevel binary observations," *Psychometrika*, vol. 65, no. 4, pp. 475–496, 2000.

[27] S.-Y. Lee and X.-Y. Song, “Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes,” *Multivariate Behavioral Research*, vol. 39, no. 4, pp. 653–686, 2004.

[28] A. Ansari, K. Jedidi, and L. Dube, “Heterogeneous factor analysis models: a Bayesian approach,” *Psychometrika*, vol. 67, no. 1, pp. 49–78, 2002.

[29] Z. Zhang, F. Hamagami, L. Wang, J. R. Nesselroade, and K. J. Grimm, “Bayesian analysis of longitudinal data using growth curve models,” *International Journal of Behavioral Development*, vol. 31, no. 4, pp. 374–383, 2007.

[30] S.-Y. Lee and X.-Y. Song, “Bayesian model selection for mixtures of structural equation models with an unknown number of components,” *The British Journal of Mathematical and Statistical Psychology*, vol. 56, no. 1, pp. 145–165, 2003.

[31] H.-T. Zhu and S.-Y. Lee, “A Bayesian analysis of finite mixtures in the LISREL model,” *Psychometrika*, vol. 66, no. 1, pp. 133–152, 2001.

[32] S.-Y. Lee, X.-Y. Song, and N.-S. Tang, “Bayesian methods for analyzing structural equation models with covariates, interaction, and quadratic latent variables,” *Structural Equation Modeling*, vol. 14, no. 3, pp. 404–434, 2007.

[33] S.-Y. Lee and N.-S. Tang, “Bayesian analysis of nonlinear structural equation models with nonignorable missing data,” *Psychometrika*, vol. 71, no. 3, pp. 541–564, 2006.

[34] G. Arminger and B. O. Muthén, “A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis-Hastings algorithm,” *Psychometrika*, vol. 63, no. 3, pp. 271–300, 1998.

[35] X.-Y. Song, S.-Y. Lee, and Y.-I. Hser, “Bayesian analysis of multivariate latent curve models with nonlinear longitudinal latent effects,” *Structural Equation Modeling*, vol. 16, no. 2, pp. 245–266, 2009.

[36] R. J. Wirth and M. C. Edwards, “Item factor analysis: current approaches and future directions,” *Psychological Methods*, vol. 12, no. 1, pp. 58–79, 2007.

[37] R. D. Bock and M. Aitkin, “Marginal maximum likelihood estimation of item parameters: application of an EM algorithm,” *Psychometrika*, vol. 46, no. 4, pp. 443–459, 1981.

[38] F. B. Baker and S.-H. Kim, Eds., *Item Response Theory: Parameter Estimation Techniques*, vol. 176 of *Statistics: Textbooks and Monographs*, Marcel Dekker, New York, NY, USA, 2nd edition, 2004.

[39] R. D. Bock and R. J. Mislevy, “Adaptive EAP estimation of ability in a microcomputer environment,” *Applied Psychological Measurement*, vol. 6, pp. 431–444, 1982.

[40] J. H. Albert, “Bayesian estimation of normal ogive item response curves using Gibbs sampling,” *Journal of Educational Statistics*, vol. 17, pp. 251–269, 1992.

[41] M. A. Tanner and W. H. Wong, “The calculation of posterior distributions by data augmentation,” *Journal of the American Statistical Association*, vol. 82, no. 398, pp. 528–550, 1987.

[42] J. H. Albert and S. Chib, “Bayesian analysis of binary and polychotomous response data,” *Journal of the American Statistical Association*, vol. 88, no. 422, pp. 669–679, 1993.

[43] S. K. Sahu, “Bayesian estimation and model choice in item response models,” *Journal of Statistical Computation and Simulation*, vol. 72, no. 3, pp. 217–232, 2002.

[44] R. J. Patz and B. W. Junker, “A straightforward approach to Markov chain Monte Carlo methods for item response models,” *Journal of Educational and Behavioral Statistics*, vol. 24, no. 2, pp. 146–178, 1999.

[45] R. J. Patz and B. W. Junker, “Applications and extensions of MCMC in IRT: multiple item types, missing data, and rated responses,” *Journal of Educational and Behavioral Statistics*, vol. 24, no. 4, pp. 342–366, 1999.

[46] J. A. Wollack, D. M. Bolt, A. S. Cohen, and Y.-S. Lee, “Recovery of item parameters in the nominal response model: a comparison of marginal maximum likelihood estimation and Markov chain Monte Carlo estimation,” *Applied Psychological Measurement*, vol. 26, no. 3, pp. 339–352, 2002.

[47] E. T. Bradlow, H. Wainer, and X. Wang, “A Bayesian random effects model for testlets,” *Psychometrika*, vol. 64, no. 2, pp. 153–168, 1999.

[48] J.-P. Fox and C. A. W. Glas, “Bayesian estimation of a multilevel IRT model using Gibbs sampling,” *Psychometrika*, vol. 66, no. 2, pp. 271–288, 2001.

[49] R. Janssen, F. Tuerlinckx, M. Meulders, and P. De Boeck, “A hierarchical IRT model for criterion-referenced measurement,” *Journal of Educational and Behavioral Statistics*, vol. 25, no. 3, pp. 285–306, 2000.

[50] Y. Takane and J. de Leeuw, “On the relationship between item response theory and factor analysis of discretized variables,” *Psychometrika*, vol. 52, no. 3, pp. 393–408, 1987.

[51] A. A. Béguin and C. A. W. Glas, “MCMC estimation and some model-fit analysis of multidimensional IRT models,” *Psychometrika*, vol. 66, no. 4, pp. 541–561, 2001.

[52] D. M. Bolt and V. F. Lall, “Estimation of compensatory and noncompensatory multidimensional item response models using Markov chain Monte Carlo,” *Applied Psychological Measurement*, vol. 27, no. 6, pp. 395–414, 2003.

[53] S. Jackman, “Multidimensional analysis of roll call data via Bayesian simulation: identification, estimation, inference, and model checking,” *Political Analysis*, vol. 9, no. 3, pp. 227–241, 2001.

[54] L. Yao and K. A. Boughton, “A multidimensional item response modeling approach for improving subscale proficiency estimation and classification,” *Applied Psychological Measurement*, vol. 31, no. 2, pp. 83–105, 2007.

[55] S.-Y. Lee and H.-T. Zhu, “Statistical analysis of nonlinear structural equation models with continuous and polytomous data,” *British Journal of Mathematical and Statistical Psychology*, vol. 53, no. 2, pp. 209–232, 2000.

[56] J.-Q. Shi and S.-Y. Lee, “Bayesian sampling-based approach for factor analysis models with continuous and polytomous data,” *British Journal of Mathematical and Statistical Psychology*, vol. 51, no. 2, pp. 233–252, 1998.

[57] X.-Y. Song and S.-Y. Lee, “Bayesian estimation and test for factor analysis model with continuous and polytomous data in several populations,” *British Journal of Mathematical and Statistical Psychology*, vol. 54, no. 2, pp. 237–259, 2001.

[58] X.-Y. Song and S.-Y. Lee, “Analysis of structural equation model with ignorable missing continuous and polytomous data,” *Psychometrika*, vol. 67, no. 2, pp. 261–288, 2002.

[59] H. Hoijtink, “Constrained latent class analysis using the Gibbs sampler and posterior predictive *p*-values: applications to educational testing,” *Statistica Sinica*, vol. 8, no. 3, pp. 691–711, 1998.

[60] H. Hoijtink and A. Notenboom, “Model based clustering of large data sets: tracing the development of spelling ability,” *Psychometrika*, vol. 69, no. 3, pp. 481–498, 2004.

[61] H. Chung, B. P. Flaherty, and J. L. Schafer, “Latent class logistic regression: application to marijuana use and attitudes among high school seniors,” *Journal of the Royal Statistical Society. Series A*, vol. 169, no. 4, pp. 723–743, 2006.

[62] M. J. H. van Onna, “Bayesian estimation and model selection in ordered latent class models for polytomous items,” *Psychometrika*, vol. 67, no. 4, pp. 519–538, 2002.

[63] A. A. Rupp and J. L. Templin, “Unique characteristics of diagnostic classification models: a comprehensive review of the current state-of-the-art,” *Measurement: Interdisciplinary Research and Perspectives*, vol. 6, no. 4, pp. 219–262, 2008.

[64] J. de la Torre and J. A. Douglas, “Higher-order latent trait models for cognitive diagnosis,” *Psychometrika*, vol. 69, no. 3, pp. 333–353, 2004.

[65] S. M. Hartz, *A Bayesian framework for the unified model for assessing cognitive abilities: blending theory with practicality*, Doctoral dissertation, Department of Statistics, University of Illinois at Urbana-Champaign, Urbana, Ill, USA, 2002.

[66] R. A. Henson, J. L. Templin, and F. Porch, “Description of the underlying algorithm of the improved Arpeggio,” Unpublished ETS Project Report, Princeton, NJ, USA, 2004.

[67] J. L. Templin and R. A. Henson, “Measurement of psychological disorders using cognitive diagnosis models,” *Psychological Methods*, vol. 11, no. 3, pp. 287–305, 2006.

[68] R. G. Almond, L. V. DiBello, B. Moulder, and J.-D. Zapata-Rivera, “Modeling diagnostic assessments with Bayesian networks,” *Journal of Educational Measurement*, vol. 44, no. 4, pp. 341–359, 2007.

[69] R. Levy and R. J. Mislevy, “Specifying and refining a measurement model for a simulation-based assessment,” *International Journal of Measurement*, vol. 4, pp. 333–369, 2004.

[70] A. A. Rupp, “Feature selection for choosing and assembling measurement models: a building-block-based organization,” *International Journal of Testing*, vol. 2, no. 3-4, pp. 311–360, 2002.

[71] D. M. Bolt, A. S. Cohen, and J. A. Wollack, “A mixture item response model for multiple-choice data,” *Journal of Educational and Behavioral Statistics*, vol. 26, no. 4, pp. 381–409, 2001.

[72] A. S. Cohen and D. M. Bolt, “A mixture model analysis of differential item functioning,” *Journal of Educational Measurement*, vol. 42, no. 2, pp. 133–148, 2005.

[73] K. A. Bollen, “Latent variables in psychology and the social sciences,” *Annual Review of Psychology*, vol. 53, pp. 605–634, 2002.

[74] S. Jackman, “Estimation and inference are missing data problems: unifying social science statistics via Bayesian simulation,” *Political Analysis*, vol. 8, pp. 307–332, 2000.

[75] L. T. Mariano and B. W. Junker, “Covariates of the rating process in hierarchical models for multiple ratings of test items,” *Journal of Educational and Behavioral Statistics*, vol. 32, no. 3, pp. 287–314, 2007.

[76] A. E. Beaton, “The NAEP 1983-1984 technical report,” Tech. Rep., ETS, Princeton, NJ, USA, 1987.

[77] S. L. Scott and E. H. Ip, “Empirical Bayes and item-clustering effects in a latent variable hierarchical model: a case study from the National Assessment of Educational Progress,” *Journal of the American Statistical Association*, vol. 97, no. 458, pp. 409–419, 2002.

[78] N. T. Longford, “Model-based methods for analysis of data from the 1990 NAEP trial state assessment,” Research and Development Report 95-696, National Center for Education Statistics, Washington, DC, USA, 1995.

[79] S. W. Raudenbush, R. P. Fotiu, and Y. F. Cheong, “Synthesizing results from the trial state assessment,” *Journal of Educational and Behavioral Statistics*, vol. 24, no. 4, pp. 413–438, 1999.

[80] M. S. Johnson and F. Jenkins, “A Bayesian hierarchical model for large-scale educational surveys: an application to the National Assessment of Educational Progress,” Research Report RR-04-38, ETS, Princeton, NJ, USA, 2005.

[81] M. S. Johnson, “A Bayesian hierarchical model for multidimensional performance assessments,” in *Proceedings of the Annual Meeting of the National Council on Measurement in Education*, New Orleans, La, USA, June 2002.

[82] R. J. Mislevy and R. D. Bock, *BILOG 3*, Scientific Software, Mooresville, Ind, USA, 2nd edition, 1990.

[83] L. K. Muthén and B. O. Muthén, *Mplus User's Guide*, Muthén & Muthén, Los Angeles, Calif, USA, 4th edition, 1998–2006.

[84] F. B. Baker, “An investigation of the item parameter recovery characteristics of a Gibbs sampling procedure,” *Applied Psychological Measurement*, vol. 22, no. 2, pp. 153–169, 1998.

[85] S. Schilling and R. D. Bock, “High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature,” *Psychometrika*, vol. 70, no. 3, pp. 533–555, 2005.

[86] B. O. Muthén, “Beyond SEM: general latent variable modeling,” *Behaviormetrika*, vol. 29, no. 1, pp. 81–117, 2002.

[87] S. Rabe-Hesketh, A. Skrondal, and A. Pickles, “Generalized multilevel structural equation modeling,” *Psychometrika*, vol. 69, no. 2, pp. 167–190, 2004.

[88] S. Rabe-Hesketh, A. Skrondal, and A. Pickles, *GLLAMM Manual*, U.C. Berkeley Division of Biostatistics Working Paper Series, 160, U.C. Berkeley Division of Biostatistics, Berkeley, Calif, USA, 2nd edition, 2004.

[89] E. Segawa, S. Emery, and S. J. Curry, “Extended generalized linear latent and mixed model,” *Journal of Educational and Behavioral Statistics*, vol. 33, no. 4, pp. 464–484, 2008.

[90] J. L. Arbuckle, *Amos 7.0 User's Guide*, SPSS, Chicago, Ill, USA, 2006.

[91] S. Jackman, “pscl: classes and methods for R developed in the political science computational laboratory,” Department of Political Science, Stanford University. Stanford, Calif, USA. R package version 1.03, 2008, http://pscl.stanford.edu/.

[92] A. D. Martin, K. M. Quinn, and J. H. Park, “pscl: MCMCpack. R package version 0.0-6,” 2009, http://cran.r-project.org/web/packages/MCMCpack/index.html.

[93] Y. Sheng, “A MATLAB package for Markov chain Monte Carlo with a multi-unidimensional IRT model,” *Journal of Statistical Software*, vol. 28, no. 10, pp. 1–20, 2008.

[94] Y. Sheng, “Markov Chain Monte Carlo estimation of normal ogive IRT models in MATLAB,” *Journal of Statistical Software*, vol. 25, no. 8, pp. 1–15, 2008.

[95] L. Yao, *BMIRT: Bayesian Multivariate Item Response Theory*, CTB/McGraw-Hill, Monterey, Calif, USA, 2003.

[96] P. Congdon, *Bayesian Statistical Modelling*, Wiley Series in Probability and Statistics, John Wiley & Sons, Chichester, UK, 2nd edition, 2006.

[97] J. Gill, *Bayesian Methods: A Social and Behavioral Sciences Approach*, Chapman & Hall/CRC, New York, NY, USA, 2nd edition, 2007.

[98] S. Sturtz, U. Ligges, and A. Gelman, “R2WinBUGS: a package for running WinBUGS from R,” *Journal of Statistical Software*, vol. 12, pp. 1–16, 2005.

[99] M. Plummer, N. Best, K. Cowles, and K. Vines, “coda: output analysis and diagnostics for MCMC,” R package version 0.13-4, 2009, http://cran.r-project.org/web/packages/coda/index.html.

[100] B. J. Smith, “boa: an R package for MCMC output convergence assessment and posterior inference,” *Journal of Statistical Software*, vol. 21, no. 11, pp. 1–37, 2007.

[101] S. Jackman, "Estimation and inference via Bayesian simulation: an introduction to Markov chain Monte Carlo," *American Journal of Political Science*, vol. 44, no. 2, pp. 375–404, 2000.

[102] J.-S. Kim and D. M. Bolt, "Estimating item response theory models using markov chain Monte Carlo methods: An NCME instructional module on," *Educational Measurement: Issues and Practice*, vol. 26, no. 4, pp. 38–51, 2007.

[103] S. Sinharay, "Experiences with Markov chain Monte Carlo convergence assessment in two psychometric examples," *Journal of Educational and Behavioral Statistics*, vol. 29, no. 4, pp. 461–488, 2004.

[104] H. Swaminathan, R. K. Hambleton, and H. J. Rogers, "Assessing the fit of item response models," in *Handbook of Statistics*, C. R. Rao and S. Sinharay, Eds., vol. 26, pp. 683–718, Elsevier/North-Holland, Amsterdam, The Netherlands, 2007.

[105] C. A. W. Glas and J. C. S. Falcón, "A comparison of item-fit statistics for the three-parameter logistic model," *Applied Psychological Measurement*, vol. 27, no. 2, pp. 87–106, 2003.

[106] D. Sörbom, "Model modification," *Psychometrika*, vol. 54, no. 3, pp. 371–384, 1989.

[107] L. Hu and P. M. Bentler, "Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives," *Structural Equation Modeling*, vol. 6, no. 1, pp. 1–55, 1999.

[108] W.-H. Chen and D. Thissen, "Local dependence indexes for item pairs using item response theory," *Journal of Educational and Behavioral Statistics*, vol. 22, no. 3, pp. 265–289, 1997.

[109] A. Gelman, X.-L. Meng, and H. Stern, "Posterior predictive assessment of model fitness via realized discrepancies," *Statistica Sinica*, vol. 6, no. 4, pp. 733–807, 1996.

[110] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde, "Bayesian measures of model complexity and fit," *Journal of the Royal Statistical Society. Series B*, vol. 64, no. 4, pp. 583–639, 2002.

[111] J. Fu, D. M. Bolt, and Y. Li, "Evaluating item fit for a polytomous fusion model using posterior predictive checks," in *Proceedings of the Annual Meeting of the National Council on Measurement in Education*, Montréal, Canada, April 2005.

[112] R. Levy, R. J. Mislevy, and S. Sinharay, "Posterior predictive model checking for multidimensionality in item response theory," *Applied Psychological Measurement*, vol. 33, no. 7, pp. 519–537, 2009.

[113] Y. Sheng and C. K. Wikle, "Comparing multiunidimensional and unidimensional item response theory models," *Educational and Psychological Measurement*, vol. 67, no. 6, pp. 899–919, 2007.

[114] S. Sinharay, "Assessing fit of unidimensional item response theory models using a Bayesian approach," *Journal of Educational Measurement*, vol. 42, no. 4, pp. 375–394, 2005.

[115] S. Sinharay, M. S. Johnson, and H. S. Stern, "Posterior predictive assessment of item response theory models," *Applied Psychological Measurement*, vol. 30, no. 4, pp. 298–321, 2006.

[116] M. J. Bayarri and J. O. Berger, "p values for composite null models," *Journal of the American Statistical Association*, vol. 95, no. 452, pp. 1127–1142, 2000.

[117] J. M. Robins, A. van der Vaart, and V. Ventura, "Asymptotic distribution of p values in composite null models," *Journal of the American Statistical Association*, vol. 95, no. 452, pp. 1143–1172, 2000.

[118] M. J. Bayarri and J. O. Berger, "Rejoinder," *Journal of the American Statistical Association*, vol. 95, pp. 1168–1170, 2000.

[119] A. Gelman, "Comment: Bayesian checking of the second levels of hierarchical models," *Statistical Science*, vol. 22, no. 3, pp. 349–352, 2007.

[120] V. E. Johnson, "Bayesian model assessment using pivotal quantities," *Bayesian Analysis*, vol. 2, no. 4, pp. 719–733, 2007.

[121] N. L. Hjort, F. A. Dahl, and G. H. Steinbakk, "Post-processing posterior predictive p-values," *Journal of the American Statistical Association*, vol. 101, no. 475, pp. 1157–1174, 2006.

[122] A. Gelman, "A Bayesian formulation of exploratory data analysis and goodness-of-fit testing," *International Statistical Review*, vol. 71, no. 2, pp. 369–382, 2003.

[123] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, The Morgan Kaufmann Series in Representation and Reasoning, Morgan Kaufmann, San Mateo, Calif, USA, 1988.