CHAPTER 38

Bayesian Structural Equation Modeling

David Kaplan
Sarah Depaoli

The history of structural equation modeling (SEM) can be roughly divided into two generations. The *first generation* of structural equation modeling began with the initial merging of confirmatory factor analysis (CFA) and simultaneous equation modeling (see, e.g., Jöreskog, 1973). In addition to these founding concepts, the first generation of SEM witnessed important methodological developments in handling nonstandard conditions of the data. These developments included methods for dealing with non-normal data, missing data, and sample size sensitivity problems (see, e.g., Kaplan, 2009). The *second generation* of SEM could be broadly characterized by another merger; this time, combining models for continuous latent variables developed in the first generation with models for categorical latent variables (see Muthén, 2001). The integration of continuous and categorical latent variables into a general modeling framework was due to the extension of finite mixture modeling to the SEM framework. This extension has provided an elegant theory, resulting in a marked increase in important applications. These applications include, but are not limited to, methods for handling the evaluation of interventions with noncompliance (Jo & Muthén, 2001), discrete-time mixture survival models (Muthén & Masyn, 2005), and models for examining unique trajectories of growth in academic outcomes (Kaplan, 2003). A more comprehensive review of the history of SEM can be found in Matsueda (Chapter 2, this volume).

A parallel development to first- and second-generation SEM has been the expansion of Bayesian methods for complex statistical models, including structural equation models. Early papers include Lee (1981), Martin and McDonald (1975), and Scheines, Hoijtink, and Boomsma (1999). A recent book by Lee (2007) provides an up-to-date review and extensions of Bayesian SEM. Most recently, B. Muthén and Asparouhov (in press) demonstrate the wide range of modeling flexibility within Bayesian SEM. The increased use of Bayesian tools for statistical modeling has come about primarily as a result of progress in computational algorithms based on Markov chain Monte Carlo (MCMC) sampling. The MCMC algorithm is implemented in software programs such as WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000), various packages within the R archive (R Development Core Team, 2008), and most recently Mplus (Muthén & Muthén, 2010).

The purpose of this chapter is to provide an accessible introduction to Bayesian SEM as an important alternative to conventional frequentist approaches to SEM. However, to fully realize the utility of the Bayesian approach to SEM, it is necessary to demonstrate not only its applicability to first-generation SEM but also how Bayesian methodology can be applied to models char-

acterizing the second generation of SEM. Although examples of Bayesian SEM relevant to first- and second-generation models will be provided, an important goal of this chapter is to develop the argument that MCMC is not just another estimation approach to SEM, but that Bayesian methodology provides a coherent philosophical alternative to conventional SEM practice, regardless of whether models are "first" or "second" generation.

The organization of this chapter is as follows. To begin, the previous chapters in this volume provide a full account of basic and advanced concepts in both first- and second-generation SEM, and we assume that the reader is familiar with these topics. Given that assumption, the next section provides a brief introduction to Bayesian ideas, including Bayes' theorem, the nature of prior distributions, description of the posterior distribution, and Bayesian model building. Following that, we provide a brief overview of MCMC sampling that we use for the empirical examples in this chapter. Next, we introduce the general form of the Bayesian structural equation model. This is followed by three examples that demonstrate the applicability of Bayesian SEM: Bayesian CFA, Bayesian multilevel path analysis, and Bayesian growth mixture modeling. Each example uses the MCMC sampling algorithm in Mplus (Muthén & Muthén, 2010). The chapter closes with a general discussion of how the Bayesian approach to SEM can lead to a pragmatic and evolutionary development of knowledge in the social and behavioral sciences.

BRIEF OVERVIEW OF BAYESIAN STATISTICAL INFERENCE

The goal of this section is to briefly present basic ideas in Bayesian inference to set the framework for Bayesian SEM, and follows closely the recent overview by Kaplan and Depaoli (in press). A good introductory treatment of the subject can be found in Hoff (2009).

To begin, denote by Y a random variable that takes on a realized value y. For example, a person's socioeconomic status could be considered a random variable taking on a very large set of possible values. In the context of SEM, Y could be vector-valued, such as items on an attitude survey. Once the person responds to the survey items, Y becomes realized as y. In a sense, Y is unobserved—it is the probability distribution of Y that we wish to understand from the actual data values y.

Next, denote by θ a parameter that we believe characterizes the probability model of interest. The parameter θ can be a scalar, such as the mean or the variance of a distribution, or it can be vector valued, such as the set of all structural model parameters, which later in the chapter we denote using the boldface $\boldsymbol{\theta}$.

We are concerned with determining the probability of observing y given unknown parameters θ, which we write as $p(y|\theta)$. In statistical inference, the goal is to obtain estimates of the unknown parameters given the data. This is expressed as the likelihood of the parameters given the data, denoted as $L(\theta|y)$. Often we work with the log-likelihood, written as $l(\theta|y)$.

The key difference between Bayesian statistical inference and frequentist statistical inference concerns the nature of the unknown parameters θ. In the frequentist tradition, the assumption is that θ is unknown but fixed. In Bayesian statistical inference, θ is random, possessing a probability distribution that reflects our uncertainty about the true value of θ. Because both the observed data y and the parameters θ are assumed random, we can model the joint probability of the parameters and the data as a function of the conditional distribution of the data given the parameters, and the prior distribution of the parameters. More formally,

$$p(\theta, y) = p(y|\theta)p(\theta) \quad \textbf{(38.1)}$$

Because of the symmetry of joint probabilities,

$$p(y|\theta)p(\theta) = p(\theta|y)p(y) \quad \textbf{(38.2)}$$

Therefore,

$$p(\theta \mid y) = \frac{p(\theta, y)}{p(y)} = \frac{p(y \mid \theta)p(\theta)}{p(y)} \quad \textbf{(38.3)}$$

where $p(\theta|y)$ is referred to as the *posterior distribution* of the parameters θ given the observed data y. Thus, from Equation 38.3, the posterior distribution of θ given y is equal to the data distribution $p(y|\theta)$ times the prior distribution of the parameters $p(\theta)$ normalized by $p(y)$ so that the distribution integrates to one. Equation 38.3 is *Bayes' theorem*. For discrete variables

$$p(y) = \sum_{\theta} p(y \mid \theta)p(\theta) \quad \textbf{(38.4)}$$

and for continuous variables

$$p(y) = \int_{\theta} p(y \mid \theta)p(\theta)d\theta \quad \textbf{(38.5)}$$

As earlier, the denominator in Equation 38.3 does not involve model parameters, so we can omit the term and obtain the *unnormalized posterior distribution*

$$p(\theta|y) \propto p(y|\theta)p(\theta) \quad \textbf{(38.6)}$$

Consider the data distribution $p(y|\theta)$ on the right hand side of Equation 38.6. When expressed in terms of the unknown parameters θ for fixed values of y, this term is the *likelihood* $L(\theta|y)$, which we mentioned earlier. Thus, Equation 38.6 can be rewritten as

$$p(\theta|y) \propto L(\theta|y)p(\theta) \quad \textbf{(38.7)}$$

Equation 38.6 represents the core of Bayesian statistical inference and is what separates Bayesian statistics from frequentist statistics. Specifically, Equation 38.6 states that our uncertainty regarding the parameters of our model, as expressed by the prior distribution $p(\theta)$, is *weighted* by the actual data $p(y|\theta)$ (or equivalently, $L[\theta|y]$), yielding an updated estimate of the model parameters, as expressed in the posterior distribution $p(\theta|y)$.

Types of Priors

The distinguishing feature of Bayesian inference is the specification of the prior distribution for the model parameters. The difficulty arises in how a researcher goes about choosing prior distributions for the model parameters. We can distinguish between two types of priors, (1) *noninformative* and (2) *informative priors,* based on how much information we believe we have prior to data collection and how accurate we believe that information to be.

Noninformative Priors

In some cases we may not be in possession of enough prior information to aid in drawing posterior inferences. From a Bayesian perspective, this lack of information is still important to consider and incorporate into our statistical specifications. In other words, it is equally as important to quantify our ignorance as it is to quantify our cumulative understanding of a problem at hand.

The standard approach to quantifying our ignorance is to incorporate a noninformative prior into our specification. Noninformative priors are also referred to as "vague" or "diffuse" priors. Arguably, the most common noninformative prior distribution is the uniform distribution over some sensible range of values. Care must be taken in the choice of the range of values over the uniform distribution. Specifically, a uniform [$-\infty$, ∞] would be an improper prior distribution insofar as it does not integrate to 1.0 as required of probability distributions. Another type of noninformative prior is the so-called "Jeffreys' prior," which handles some of the problems associated with uniform priors. An important treatment of noninformative priors can be found in Press (2003).

Informative Priors

In many practical situations, there may be sufficient prior information on the shape and scale of the distribution of a model parameter that it can be systematically incorporated into the prior distribution. Such priors are referred to as "informative." One type of informative prior is based on the notion of a "conjugate prior" distribution, which is one that, when combined with the likelihood function, yields a posterior distribution that is in the same distributional family as the prior distribution. This is a very important and convenient feature because if a prior is not conjugate, the resulting posterior distribution may have a form that is not analytically simple to solve. Arguably, the existence of numerical simulation methods for Bayesian inference, such as MCMC sampling, may render nonconjugacy less of a problem.

Point Estimates of the Posterior Distribution

Bayes' theorem shows that the posterior distribution is composed of encoded prior information weighted by the data. With the posterior distribution in hand, it is of interest to obtain summaries of the distribution—such as the mean, mode, and variance. In addition, interval summaries of the posterior distribution can be obtained. Summarizing the posterior distribution provides the necessary ingredients for Bayesian hypothesis testing. In the general case, the expressions for the mean and variance of the posterior distribution come from expressions for the mean and variance of conditional distributions generally. Specifically, for the continuous case, the mean of the posterior distribution can be written as

$$E(\theta \mid y) = \int_{-\infty}^{+\infty} \theta p(\theta \mid y) d\theta \quad \textbf{(38.8)}$$

and is referred to as the *expected a posteriori* or EAP estimate. Thus, the conditional expectation of θ is obtained by averaging over the marginal distribution of *y*. Similarly, the conditional variance of θ can be obtained as (see Gill, 2002)

$$\begin{aligned} var(\theta \mid y) &= E[(\theta - E[(\theta \mid y])^2 \mid y) \\ &= E(\theta^2 \mid y) - E(\theta \mid y)^2 \end{aligned} \quad \textbf{(38.9)}$$

The conditional expectation and variance of the posterior distribution provide two simple summary values of the distribution. Another summary measure would be the mode of the posterior distribution. Those measures, along with the quantiles of the posterior distribution, provide a complete description of the distribution.

Credibility Intervals

One important consequence of viewing parameters probabilistically concerns the interpretation of "confidence intervals." Recall that the frequentist confidence interval is based on the assumption of a very large number of repeated samples from the population characterized by a fixed and unknown parameter μ. For any given sample, we obtain the sample mean $\bar{x}$ and form, for example, a 95% confidence interval. The correct frequentist interpretation is that 95% of the confidence intervals formed this way capture the true parameter μ under the null hypothesis. Notice that from this perspective, the probability that the parameter is in the interval is either zero or one.

In contrast, the Bayesian perspective forms a "credibility interval" (also known as a "posterior probability interval"). Again, because we assume that a parameter has a probability distribution, when we sample from the posterior distribution of the model parameters, we can obtain its quantiles. From the quantiles, we can directly obtain the probability that a parameter lies within a particular interval. So in this example, a 95% credibility interval means that the probability that the parameter lies in the interval is 0.95. Notice that this is entirely different from the frequentist interpretation, and arguably aligns with common sense.

Formally, a $100(1 - \alpha)\%$ credibility interval for a particular subset of the parameter space θ is defined as

$$1 - \alpha = \int_C p(\theta \mid x) d\theta \quad \textbf{(38.10)}$$

Highest Posterior Density

The simplicity of the credibility interval notwithstanding, it is not the only way to provide an interval estimate of a parameter. Following the argument set down by Box and Tiao (1973), when considering the posterior distribution of a parameter θ, there is a substantial part of the region of that distribution where the density is quite small. It may be reasonable, therefore, to construct an interval in which every point inside has a higher probability than any point outside the interval. Such a construction is referred to as the *highest probability density* (HPD) interval. More formally,

Definition 1

Let $p(\theta \mid y)$ *be the posterior probability density function. A region R of the parameter space* θ *is called the HPD region of the interval* $1 - \alpha$ *if*

1. $P(\theta \in R \mid y) = 1 - \alpha$
2. *For* $\theta_1 \in R$ *and* $\theta_2 \notin R$, $p(\theta_1 \mid y) \geq p(\theta_2 \mid y)$.

In words, the first part says that given the data *y*, the probability is that θ is in a particular region defined as $1 - \alpha$, where α is determined ahead of time. The second part says that for two different values of θ, denoted as θ_1 and θ_2, if θ_1 is in the region defined by $1 - \alpha$, but θ_2 is not, then θ_1 has a higher probability than θ_2 given the data. Note that for unimodal and symmetric distributions, such as the uniform distribution or the normal distribution, the HPD is formed by choosing tails of equal density. The advantage of the HPD arises when densities are not symmetric and/or are not unimodal. In fact, this is an important property of the HPD and sets it apart from standard credibility intervals. Following Box and Tiao (1973), if $p(\theta \mid y)$ is not uniform over every region in θ, then the HPD region $1 - \alpha$ is unique. Also if $p(\theta_1 \mid y) = p(\theta_2 \mid y)$, then these points are included (or excluded) by a $1 - \alpha$ HPD region. The opposite is true as well, namely, if $p(\theta_1 \mid y) \neq p(\theta_2 \mid y)$, then a $1 - \alpha$ HPD region includes one point but not the other (Box & Tiao, 1973, p. 123).

BAYESIAN MODEL EVALUATION AND COMPARISON

SEM, by its very nature, involves the specification, estimation, and testing of models that purport to represent the underlying structure of data. In this case, SEM is

not only a noun describing a broad class of methodologies, but it is also a verb—an activity on the part of a researcher to describe and analyze a phenomenon of interest. The chapters in this handbook have described the nuances of SEM from the frequentist domain—with many authors attending to issues of specification, power, and model modification. In this section, we consider model evaluation and comparison from the Bayesian perspective. We focus on two procedures that are available in Mplus, namely, posterior predictive checking along with posterior predictive *p*-values as a means of evaluating the quality of the fit of the model (see, e.g., Gelman, Carlin, Stern, & Rubin, 2003), and the deviance information criterion for the purposes of model comparison (Spiegelhalter, Best, Carlin, & van der Linde, 2002). We are quick to note, however, that these procedures are available in WinBUGS as well as various programs within the R environment such as LearnBayes (Albert, 2007) and MCMCpack (Martin, Quinn, & Park, 2010).

Posterior Predictive Checks

The general idea behind posterior predictive checking is that there should be little, if any, discrepancy between data generated by the model, and the actual data itself. In essence, posterior predictive checking is a method for assessing the specification quality of the model from the viewpoint of predictive accuracy. Any deviation between the model-generated data and the actual data suggests possible model misspecification.

Posterior predictive checking utilizes the posterior predictive distribution of replicated data. Following Gelman and colleagues (2003), let y^{rep} be data replicated from our current model. That is,

$$p(y^{\text{rep}} \mid y) = \int p(y^{\text{rep}} \mid \theta) p(\theta \mid y) d\theta \qquad \textbf{(38.11)}$$
$$= \int p(y^{\text{rep}} \mid \theta) p(y \mid \theta) p(\theta) d\theta$$

Notice that the second term, $p(\theta|y)$, on the right-hand side of Equation 38.11 is simply the posterior distribution of the model parameters. In words, Equation 38.11 states that the distribution of future observations given the present data, $p(y^{\text{rep}}|y)$, is equal to the probability distribution of the future observations given the parameters, $p(y^{\text{rep}}|\theta)$, weighted by the posterior distribution of the model parameters. Thus, posterior predictive checking accounts for both the uncertainty in the model parameters and the uncertainty in the data.

As a means of assessing the fit of the model, posterior predictive checking implies that the replicated data should match the observed data quite closely if we are to conclude that the model fits the data. One approach to quantifying model fit in the context of posterior predictive checking incorporates the notion of Bayesian *p*-values. Denote by $T(y)$ a model test statistic based on the data, and let $T(y^{\text{rep}})$ be the same test statistic but defined for the replicated data. Then, the Bayesian *p*-value is defined to be

$$p\text{-value} = pr(T(y^{\text{rep}}) \geq T(y) \mid y) \qquad \textbf{(38.12)}$$

Equation 38.12 measures the proportion of test statistics in the replicated data that exceeds that of the actual data. We will demonstrate posterior predictive checking in our examples.

Bayes Factors

As suggested earlier in this chapter, the Bayesian framework does not adopt the frequentist orientation to null hypothesis significance testing. Instead, as with posterior predictive checking, a key component of Bayesian statistical modeling is a framework for model choice, with the idea that the model will be used for prediction. For this chapter, we will focus on Bayes factors, the Bayesian information criterion, and the deviance information criterion as methods for choosing among a set of competing models. The deviance information criterion will be used in the subsequent empirical examples.

A very simple and intuitive approach to model building and model selection uses so-called "Bayes factors" (Kass & Raftery, 1995). An excellent discussion of Bayes factors and the problem of hypothesis testing from the Bayesian perspective can be found in Raftery (1995). In essence, the Bayes factor provides a way to quantify the odds that the data favor one hypothesis over another. A key benefit of Bayes factors is that models do not have to be nested.

To begin, consider two competing models, denoted as M_1 and M_2, that could be nested within a larger space of alternative models. For example, these could be two regression models with a different number of variables, or two structural equation models specifying very different directions of mediating effects. Further, let θ_1 and θ_2 be two parameter vectors. From Bayes' theorem, the posterior probability that, say, M_1, is the correct model can be written as

$$p(M_1 \mid y) = \frac{p(y \mid M_1)p(M_1)}{p(y \mid M_1)p(M_1) + p(y \mid M_2)p(M_2)} \quad \textbf{(38.13)}$$

Notice that $p(y|M_1)$ does not contain model parameters θ_1. To obtain $p(y|M_1)$ requires integrating over θ_1. That is

$$p(y \mid M_1) = \int p(y \mid \theta_1, M_1)p(\theta_1 \mid M_1)d\theta_1 \quad \textbf{(38.14)}$$

where the terms inside the integral are the likelihood and the prior, respectively. The quantity $p(y|M_1)$ has been referred to as the "integrated likelihood" for model M_1 (Raftery, 1995). Perhaps a more useful term is the "predictive probability of the data" given M_1. A similar expression can be written for M_2.

With these expressions, we can move to the comparison of our two models, M_1 and M_2. The goal is to develop a quantity that expresses the extent to which the data support M_1 over M_2. One quantity could be the posterior odds of M_1 over M_2, expressed as

$$\frac{p(M_1 \mid y)}{p(M_2 \mid y)} = \frac{p(y \mid M_1)}{p(y \mid M_2)} \times \left[\frac{p(M_1)}{p(M_2)}\right] \quad \textbf{(38.15)}$$

Notice that the first term on the right-hand side of Equation 38.15 is the ratio of two integrated likelihoods. This ratio is referred to as the "Bayes factor" for M_1 over M_2, denoted here as B_{12}. In line with Kass and Raftery (1995, p. 776), our prior opinion regarding the odds of M_1 over M_2, given by $p(M_1)/p(M_2)$, is weighted by our consideration of the data, given by $p(y|M_1)/p(y|M_2)$. This weighting gives rise to our updated view of evidence provided by the data for either hypothesis, denoted as $p(M_1|y)/p(M_2|y)$. An inspection of Equation 38.15 also suggests that the Bayes factor is the ratio of the posterior odds to the prior odds.

In practice, there may be no prior preference for one model over the other. In this case, the prior odds are neutral and $p(M_1) = p(M_2) = 1/2$. When the prior odds ratio equals 1, then the posterior odds is equal to the Bayes factor.

The Bayesian Information Criterion

A popular measure for model selection used in both frequentist and Bayesian applications is based on an approximation of the Bayes factor and is referred to as the "Bayesian information criterion" (BIC), also called the "Schwarz criterion" (Schwarz, 1978). A detailed mathematical derivation for the BIC can be found in Raftery (1995), who also examines generalizations of the BIC to a broad class of statistical models.

Under conditions where there is little prior information, Raftery (1995) has shown that an approximation of the Bayes factor can be written as

$$\text{BIC} = -2\log(\hat{\theta} \mid y) + q\log(n) \quad \textbf{(38.16)}$$

where $-2\log(\hat{\theta}|y)$ describes model fit, while $q\log(n)$ is a penalty for model complexity, q represents the number of variables in the model, and n is the sample size.

As with Bayes factors, the BIC is often used for model comparisons. Specifically, the difference between two BIC measures comparing, say, M_1 to M_2 can be written as

$$\begin{aligned}\Delta(\text{BIC}_{12}) &= \text{BIC}_{(M_1)} - \text{BIC}_{(M_2)} \\ &= \log(\hat{\theta}_1 \mid y) - \log(\hat{\theta}_2 \mid y) - \frac{1}{2}(q_1 - q_2)\log(n)\end{aligned} \quad \textbf{(38.17)}$$

Rules of thumb have been developed to assess the quality of the evidence favoring one hypothesis over another using Bayes factors and the comparison of BIC values from two competing models. Following Kass and Raftery (1995, p. 777) and using M_1 as the reference model,

BIC difference	Bayes factor	Evidence against M_2
0 to 2	1 to 3	Weak
2 to 6	3 to 20	Positive
6 to 10	20 to 150	Strong
> 10	> 150	Very strong

The Deviance Information Criterion (DIC)

Although the BIC is derived from a fundamentally Bayesian perspective, it is often productively used for model comparison in the frequentist domain. Recently, however, an explicitly Bayesian approach to model comparison was developed by Spiegelhalter and colleagues (2002) based on the notion of *Bayesian deviance*.

Consider a particular probability model for a set of data, defined as $p(y|\theta)$. Then, *Bayesian deviance* can be defined as

$$D(\theta) = -2\log[p(y \mid \theta)] + 2\log[h(y)] \quad \textbf{(38.18)}$$

where, according to Spielgelhalter and colleagues (2002), the term $h(y)$ is a standardizing factor that does not involve model parameters and thus is not involved in model selection. Note that although Equation 38.18 is similar to the BIC, it is not, as currently defined, an explicit Bayesian measure of model fit. To accomplish this, we use Equation 38.18 to obtain a posterior mean over θ by defining

$$\text{DIC} = E_{\theta}\{-2\log[p(y|\theta)|y] + 2\log[h(y)\} \quad \textbf{(38.19)}$$

Similar to the BIC, the model with the smallest DIC among a set of competing models is preferred.

BRIEF OVERVIEW OF MCMC ESTIMATION

As stated in the introduction, the key reason for the increased popularity of Bayesian methods in the social and behavioral sciences has been the advent of powerful computational algorithms now available in proprietary and open-source software. The most common algorithm for Bayesian estimation is based on MCMC sampling. A number of very important papers and books have been written about MCMC sampling (see, e.g., Gilks, Richardson, & Spiegelhalter, 1996). Suffice it to say, the general idea of MCMC is that instead of attempting to analytically solve for the moments and quantiles of the posterior distribution, MCMC instead draws specially constructed samples from the posterior distribution $p(\theta|y)$ of the model parameters.

The formal algorithm can be specified as follows. Let $\boldsymbol{\theta}$ be a vector of model parameters with elements $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)'$. Note that information regarding $\boldsymbol{\theta}$ is contained in the prior distribution $p(\boldsymbol{\theta})$. A number of algorithms and software programs are available to conduct MCMC sampling. For the purposes of this chapter, we use the Gibbs sampler (Geman & Geman, 1984) as implemented in Mplus (Muthén & Muthén, 2010). Following the description given in Hoff (2009), the Gibbs sampler begins with an initial set of starting values for the parameters, denoted as $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_q^{(0)})'$. Given this starting point, the Gibbs sampler generates $\theta^{(s)}$ from $\theta^{(s-1)}$ as follows:

$$\begin{aligned}
&1.\ \text{sample}\quad \theta_1^{(s)} \sim p(\theta_1 \mid \theta_2^{(s-1)}, \theta_3^{(s-1)}, \ldots, \theta_q^{(s-1)}, \boldsymbol{y})\\
&2.\ \text{sample}\quad \theta_2^{(s)} \sim p(\theta_2 \mid \theta_1^{(s-1)}, \theta_3^{(s-1)}, \ldots, \theta_q^{(s-1)}, \boldsymbol{y})\\
&\qquad\qquad\vdots\\
&q.\ \text{sample}\quad \theta_q^{(s)} \sim p(\theta_q \mid \theta_1^{(s)}, \theta_2^{(s)}, \ldots, \theta_{q-1}^{(s)}, \boldsymbol{y})
\end{aligned}$$

where $s = 1, 2, \ldots, S$ are the Monte Carlo interations. Then, a sequence of dependent vectors is formed

$$\begin{aligned}
\boldsymbol{\theta}^{(1)} &= \left\{\theta_1^{(1)}, \ldots, \theta_q^{(1)}\right\}\\
\boldsymbol{\theta}^{(2)} &= \left\{\theta_1^{(2)}, \ldots, \theta_q^{(2)}\right\}\\
&\vdots\\
\boldsymbol{\theta}^{(S)} &= \left\{\theta_1^{(S)}, \ldots, \theta_q^{(S)}\right\}
\end{aligned}$$

This sequence exhibits the so-called "Markov property" insofar as $\theta^{(s)}$ is conditionally independent of $\{\theta_1^{(0)}, \ldots \theta_q^{(s-2)}\}$ given $\theta^{(s-1)}$. Under some general conditions, the sampling distribution resulting from this sequence will converge to the target distribution as $S \to \infty$. See Gilks and colleagues (1996) for additional details on the properties of MCMC.

In setting up the Gibbs sampler, a decision must be made regarding the number of Markov chains to be generated, as well as the number of iterations of the sampler. With regard to the number of chains to be generated, it is not uncommon to specify multiple chains. Each chain samples from another location of the posterior distribution based on purposefully disparate starting values. With multiple chains it may be the case that fewer iterations are required, particularly if there is evidence for the chains converging to the same posterior mean for each parameter. Convergence can also be obtained from one chain, though often requiring a considerably larger number of iterations. Once the chain has stabilized, the iterations prior to the stabilization (referred to as the "burn-in" phase) are discarded. Summary statistics, including the posterior mean, mode, standard deviation and credibility intervals, are calculated on the post-burn-in iterations.[1]

Convergence Diagnostics

Assessing the convergence of parameters within MCMC estimation is a difficult task that has received considerable attention in the literature (see, e.g., Sinharay, 2004). The difficulty of assessing convergence stems from the very nature of the MCMC algorithm because it is designed to converge in distribution rather than to a point estimate. Because there is not a single adequate assessment of convergence for this situation, it is common to inspect several different diagnostics that examine varying aspects of convergence conditions.

A variety of these diagnostics are reviewed and demonstrated in Kaplan and Depaoli (in press), including the Geweke (1992) convergence diagnostic, the Heidelberger and Welch (1983) convergence diagnostic, and the Raftery and Lewis (1992) convergence diagnostic. These diagnostics can be used for the single-chain situation.

One of the most common diagnostics in a multiple-chain situation is the Brooks, Gelman, and Rubin diagnostic (see, e.g., Gelman, 1996; Gelman & Rubin, 1992a, 1992b). This diagnostic is based on analysis of variance and is intended to assess convergence among several parallel chains with varying starting values. Specifically, Gelman and Rubin (1992a) proposed a method where an overestimate and an underestimate of the variance of the target distribution are formed. The overestimate of variance is represented by the between-chain variance, and the underestimate is the within-chain variance (Gelman, 1996). The theory is that these two estimates would be approximately equal at the point of convergence. The comparison of between and within variances is referred to as the "potential scale reduction factor" (PSRF), and larger values typically indicate that the chains have not fully explored the target distribution. Specifically, a variance ratio that is computed with values approximately equal to 1.0 indicates convergence. Brooks and Gelman (1998) added an adjustment for sampling variability in the variance estimates and also proposed a multivariate extension (MPSRF), which does not include the sampling variability correction. The changes by Brooks and Gelman reflect the diagnostic as implemented in Mplus (Muthén & Muthén, 2010).

SPECIFICATION OF BAYESIAN SEM

Following general notation, denote the measurement model as

$$y = \alpha + \Lambda\eta + \mathbf{K}x + \varepsilon \qquad (38.20)$$

where $\boldsymbol{y}$ is a vector of manifest variables, α is a vector of measurement intercepts, Λ is a factor loading matrix, η is a vector of latent variables, $\mathbf{K}$ is a matrix of regression coefficients relating the manifest variables $\boldsymbol{y}$ to observed variables $\boldsymbol{x}$, and ε is a vector of uniquenesses with covariance matrix Ξ, assumed to be diagonal. The structural model relating common factors to each other and possibly to a vector of manifest variables x is written as

$$\eta = \nu + \mathbf{B}\eta + \Gamma x + \zeta \qquad (38.21)$$

where ν is a vector of structural intercepts, $\mathbf{B}$ and Γ are matrices of structural coefficients, and ζ is a vector of structural disturbances with covariance matrix Ψ, which is assumed to be diagonal.

Conjugate Priors for SEM Parameters

To specify the prior distributions, it is notationally convenient to arrange the model parameters as sets of common conjugate distributions. Parameters with the subscript 'norm' follow a normal distribution, while those with the subscript 'IW' follow an inverse-Wishart distribution. Let $\theta_{\text{norm}} = \{\alpha, \nu, \Lambda, \mathbf{B}, \Gamma, \mathbf{K}\}$ be the vector of free model parameters that are assumed to follow a normal distribution, and let $\theta_{\text{IW}} = \{\Xi, \Psi\}$ be the vector of free model parameters that are assumed to follow the inverse-Wishart distribution. Formally, we write

$$\theta_{\text{norm}} \sim N(\mu, \Omega) \qquad (38.22)$$

where μ and Ω are the mean and variance hyperparameters, respectively, of the normal prior. For blocks of variances and covariances in Ξ and Ψ, we assume that the prior distribution is IW,[2] that is,

$$\theta_{\text{IW}} \sim \text{IW}(\mathbf{R}, \delta) \qquad (38.23)$$

where $\mathbf{R}$ is a positive definite matrix, and $\delta > q - 1$, where q is the number of observed variables. Different choices for $\mathbf{R}$ and δ will yield different degrees of "informativeness" for the IW distribution.

In addition to the conventional SEM model parameters and their priors, an additional model parameter is required for the growth mixture modeling example given below. Specifically, it is required that we estimate the mixture proportions, which we denote as π. In this specification, the class labels assigning an individual to a particular trajectory class follow a multinomial distribution with parameters n, the sample size, and π is a vector of trajectory class proportions. The conjugate prior for trajectory class proportions is the Dirichlet(τ) distribution with hyperparameters $\tau = (\tau_1, \ldots, \tau_T)$, where T is the number of trajectory classes and $\sum_{T=1}^{T} = 1$.

MCMC Sampling for Bayesian SEM

The Bayesian approach begins by considering $\boldsymbol{\eta}$ as missing data. Then, the observed data y are augmented with $\boldsymbol{\eta}$ in the posterior analysis. The Gibbs sampler then produces a posterior distribution $[\boldsymbol{\theta}_n, \boldsymbol{\theta}_{\text{IW}}, \boldsymbol{\eta} | \boldsymbol{y}]$ via the following algorithm. At the $(s + 1)$th iteration, using current values of $\boldsymbol{\eta}^{(s)}$, $\boldsymbol{\theta}_{\text{norm}}^{(s)}$, and $\boldsymbol{\theta}_{\text{IW}}^{(s)}$,

1. sample $\boldsymbol{\eta}^{(s+1)}$ from $p(\boldsymbol{\eta} \,|\, \boldsymbol{\theta}_{\text{norm}}^{(s)}, \boldsymbol{\theta}_{\text{IW}}^{(s)}, \boldsymbol{y})$ **(38.24)**
2. sample $\boldsymbol{\theta}_n^{(s+1)}$ from $p(\boldsymbol{\theta}_{\text{norm}} \,|\, \boldsymbol{\theta}_{\text{IW}}^{(s)}, \boldsymbol{\eta}^{(s+1)}, \boldsymbol{y})$ **(38.25)**
3. sample $\boldsymbol{\theta}_{\text{IW}}^{(s+1)}$ from $p(\boldsymbol{\theta}_{\text{IW}} \,|\, \boldsymbol{\theta}_{\text{norm}}^{(s+1)}, \boldsymbol{\eta}^{(s+1)}, \boldsymbol{y})$ **(38.26)**

In words, Equations 38.24–38.26 first require start values for $\boldsymbol{\theta}_{\text{norm}}^{(0)}$ and $\boldsymbol{\theta}_{\text{IW}}^{(0)}$ to begin the MCMC generation. Then, given these current start values and the data $\boldsymbol{y}$ at iteration s, we generate $\boldsymbol{\eta}$ at iteration $s + 1$. Given the latent data and observed data, we generate estimates of the measurement model and structural model parameters in Equations 38.20 and 38.21, respectively. The computational details can be found in Asparouhov and Muthén (2010).

THREE EXAMPLES OF BAYESIAN SEM

This section provides three examples of Bayesian SEM. Example 1 presents a simple two-factor Bayesian CFA. This model is compared to an alternative model with only one factor. Example 2 presents an example of a multilevel path analysis with a randomly varying slope. Example 3 presents Bayesian growth mixture modeling.

Bayesian CFA

Data for this example is comprised of an unweighted sample of 665 kindergarten teachers from the fall assessment of the Early Childhood Longitudinal Study—Kindergarten (ECLS-K) class of 1998–1999 (National Center for Education Statistics [NCES], 2001). The teachers were given a questionnaire about different characteristics of the classroom and students. A portion of this questionnaire consisted of a series of Likert-type items regarding the importance of different student characteristics and classroom behavior. Nine of these items were chosen for this example. All items were scored based on a 5-point summative response scale regarding the applicability and importance of each item to the teacher.

For this example we presume to have strong prior knowledge of the factor loadings, but no prior knowledge of the factor means, factor variances, and unique variances. For the factor loadings, strong prior knowledge can be determined as a function of both the location and the precision of the prior distribution. In particular, the mean hyperparameter would reflect the prior knowledge of the factor loading value (set at 0.8 in this example), and the precision of the prior distribution would be high (small variances of 0.01 were used here) to reflect the strength of our prior knowledge. As the strength of our knowledge decreases for a parameter, the variance hyperparameter would increase to reflect our lack of precision in the prior.

For the factor means, factor variances, and unique variances, we specified priors that reflected no prior knowledge about those parameters. The factor means were given prior distributions that were normal but contained very little precision. Specifically, the mean hyperparameters were set arbitrarily at 0, and the variance hyperparameters were specified as 10^{10} to indicate no precision in the prior. The factor variances and unique variances also received priors reflecting no prior knowledge about those parameters. These variance parameters all received IW priors that were completely diffuse, as described in Asparouhov and Muthén (2010).

On the basis of preliminary exploratory factor analyses, the CFA model in this example is specified to have two factors. The first factor contains two items related to the importance teachers place on how a student's progress relates to other children. The items specifically address how a student's achievements compare to other students in the classroom and also how they compare to statewide standards. The second factor comprises seven items that relate to individual characteristics of the student. These items include the following topics: improvement over past performance, overall effort, class participation, daily attendance, classroom behavior, cooperation with other students, and the ability to follow directions.

Parameter Convergence

A CFA model was estimated with 10,000 total iterations, 5,000 burn-in and 5,000 post-burn-in. This model converged properly as indicated by the Brooks

and Gelman (1998) (PSRF) diagnostic. Specifically, the estimated value for PSRF fell within a specified range surrounding 1.0. This model took less than 1 minute to compute.

Figure 38.1 presents convergence plots, posterior density plots, and autocorrelation plots (for both chains) for the factor loadings for items 2 and 4. Perhaps the most common form of assessing MCMC convergence is to examine the convergence (also called "history") plots produced for a chain. Typically, a parameter will appear to converge if the sample estimates form a tight horizontal band across this history plot. This method is more likely to be an indicator of nonconvergence. It is typical to use multiple Markov chains, each with different starting values, to assess parameter convergence. For example, if two separate chains for the same parameter are sampling from different areas of the target distribution, there is evidence of nonconvergence. Like-

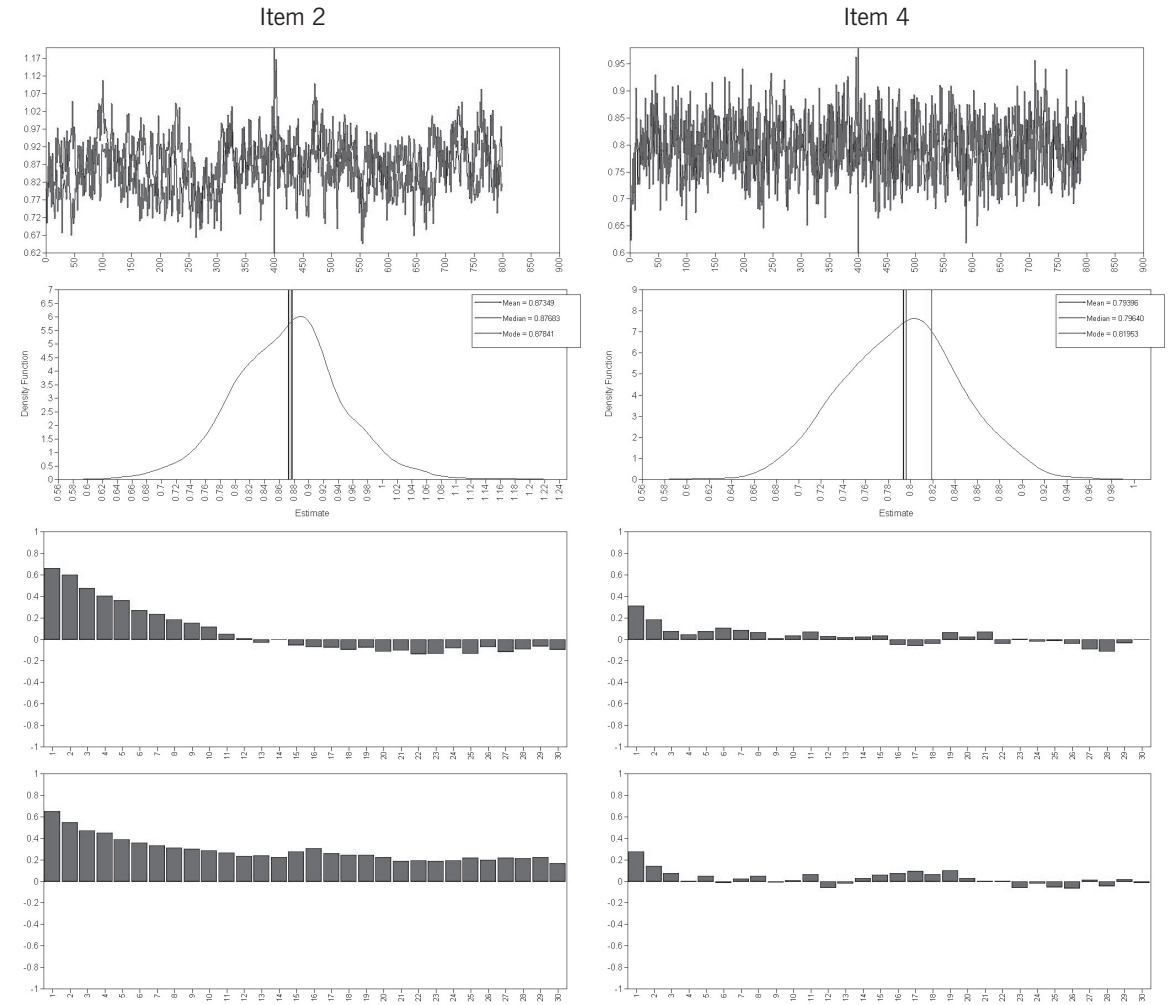

FIGURE 38.1. CFA: Convergence, posterior densities, and autocorrelation plots for select parameters.

wise, if a plot shows substantial fluctuation or jumps in the chain, it is likely the parameter has not reached convergence. The convergence plots in Figure 38.1 exhibit a tight, horizontal band for both of the parameters presented. This tight band indicates the parameters likely converged properly.

Next, Figure 38.1 presents the posterior probability density plots that indicate the posterior densities for these parameters are approximating a normal density. The following two rows present the autocorrelation plots for each of the two chains. Autocorrelation plots illustrate the amount of dependence in the chain. These plots represent the post-burn-in phase of the respective chains. Each of the two chains for these parameters shows relatively low dependence, indicating that the estimates are not being impacted by starting values or by the previous sampling states in the chain.

The other parameters included in this model showed similar results of proper convergence, normal posterior densities, and low autocorrelations for both MCMC chains. Appendix 38.1 contains the Mplus code for this example.

Model Interpretation

Estimates based on the post-burn-in iterations for the final CFA model are presented in Table 38.1. The EAP estimates and standard deviations of the posterior distributions are provided for each parameter. The one-tailed *p*-value based on the posterior distribution is also included for each parameter. If the parameter estimate is positive, this *p*-value represents the proportion of the posterior distribution that is below zero. If the parameter estimate is negative, the *p*-value is the proportion of the posterior distribution that is above zero (B. Muthén, 2010, p. 7). Finally, the 95% credibility interval is provided for each parameter. The first factor consisted of measures comparing the student's progress to others, while the second factor consisted of individual student characteristics. Note that the first item on each factor was fixed to have a loading of 1.00 in order to set the metric of that factor.

The factor comparing the student's progress to state standards has a high loading of 0.87. The factor measuring individual student characteristics also had high factor loadings, ranging from 0.79 to 1.10 (unstandardized). Note that although these are unstandardized loadings, the Bayesian estimation framework can handle any form of standardization as well. Estimates for factor variances and covariances, factor means, and residual variances are also included in Table 38.1.

The one-sided *p*-values in Table 38.1 can aid in interpreting the credibility interval produced by the posterior distribution. For example, in the case of the means for factor 1 and factor 2, the lower bound of the 95% credibility interval was negative and the upper bound was positive. The one-sided *p*-value indicates exactly what proportion of the posterior is negative and what proportion is positive. For the factor 1 mean, the *p*-value indicated that 13% of the posterior distribution fell below zero. Likewise, results for the factor 2 mean indicated that 45% of the posterior distribution fell below zero. Overall, these *p*-values, especially for the factor 2 mean, indicated that a large portion of the posterior distribution was negative even though the EAP estimate was positive.

Model Fit and Model Comparison

For this example, we illustrate posterior predictive checking (PPC) for model assessment, and the DIC for model choice. Specifically, PPC was demonstrated for the two-factor CFA model, and the DIC was used to compare the two-factor CFA model to a one-factor CFA model.

In Mplus, PPC uses the likelihood ratio chi-square test as the discrepancy function between the actual data and the data generated by the model. A posterior predictive *p*-value is then computed based on this discrepancy function. Unlike the classical *p*-value, the Bayesian *p*-value takes into account the variability of the model parameters and does not rely on asymptotic theory (Asparouhov & Muthén, 2010, p. 28). As mentioned, the data generated by the model should closely match the observed data if the model fits. Specifically, if the posterior predictive *p*-value obtained is small, this is an indication of model misfit for the observed data. The PPC test also produces a 95% confidence interval for the difference between the value of the chi-square model test statistic for the observed sample data and that for the replicated data (Muthén, 2010).

Model fit was assessed by PPC for the original two-factor CFA model presented earlier. The model was rejected based on the PPC test with a posterior predictive *p*-value of .00, indicating that the model does not adequately represent the observed data. The 95% confidence interval for the difference between the observed data test statistic and the replicated data test statistic

TABLE 38.1. MCMC CFA Estimates: ECLS-K Teacher Survey

Parameter	EAP	*SD*	*p*-value	95% credibility interval
Loadings: Compared to others				
Compared to other children	1.00			
Compared to state standards	0.87	0.07	0.00	0.73, 1.02
Loadings: Individual characteristics				
Improvement	1.00			
Effort	0.79	0.05	0.00	0.70, 0.89
Class participation	1.09	0.06	0.00	0.97, 1.20
Daily attendance	1.08	0.06	0.00	0.96, 1.20
Class behavior	1.10	0.05	0.00	1.00, 1.20
Cooperation with others	1.10	0.05	0.00	1.00, 1.20
Follow directions	0.82	0.05	0.00	0.72, 0.91
Factor means				
Factor 1 mean	0.30	0.22	0.13	–0.07, 0.65
Factor 2 mean	0.02	0.07	0.45	–0.08, 0.18
Factor variances and covariances				
Factor 1 variance	0.45	0.05	0.00	0.35, 0.55
Factor 2 variance	0.14	0.01	0.00	0.12, 0.17
Factor covariance	0.11	0.01	0.00	0.09, 0.14
Residual variances				
Compared to other children	0.31	0.04	0.00	0.23, 0.39
Compared to state standards	0.60	0.05	0.00	0.52, 0.70
Improvement	0.28	0.02	0.00	0.25, 0.31
Effort	0.21	0.01	0.00	0.18, 0.23
Class participation	0.27	0.02	0.00	0.23, 0.30
Daily attendance	0.29	0.02	0.00	0.26, 0.33
Classroom behavior	0.16	0.01	0.00	0.13, 0.18
Cooperation with others	0.17	0.01	0.00	0.14, 0.19
Follow directions	0.18	0.01	0.00	0.16, 0.20

had a lower bound of 149.67 and an upper bound of 212.81 (see Figure 38.2). Since the confidence interval for the difference in the observed and replicated data is positive, this indicates "that the observed data test statistic is much larger than what would have been generated by the model" (Muthén, 2010, p. 14).

Figure 38.2 illustrates the PPC plot and the corresponding PPC scatterplot for the original two-factor model. The PPC distribution plot shows the distribution of the difference between the observed data test statistic and the replicated data test statistic. In this plot, the observed data test statistic is marked by the *y*-axis line, which corresponds to a value of zero on the *x*-axis. The PPC scatterplot, also presented in Figure 38.2, has a 45 degree line that helps to define the posterior predictive *p*-value. With all of the points below this line, this indicates that the *p*-value (0.00) was quite small and the model can be rejected, indicating model misfit for the observed data. If adequate model fit had been observed, the points would be plotted along the 45 degree line in Figure 38.2, which would indicate a close match between the observed and the replicated data.

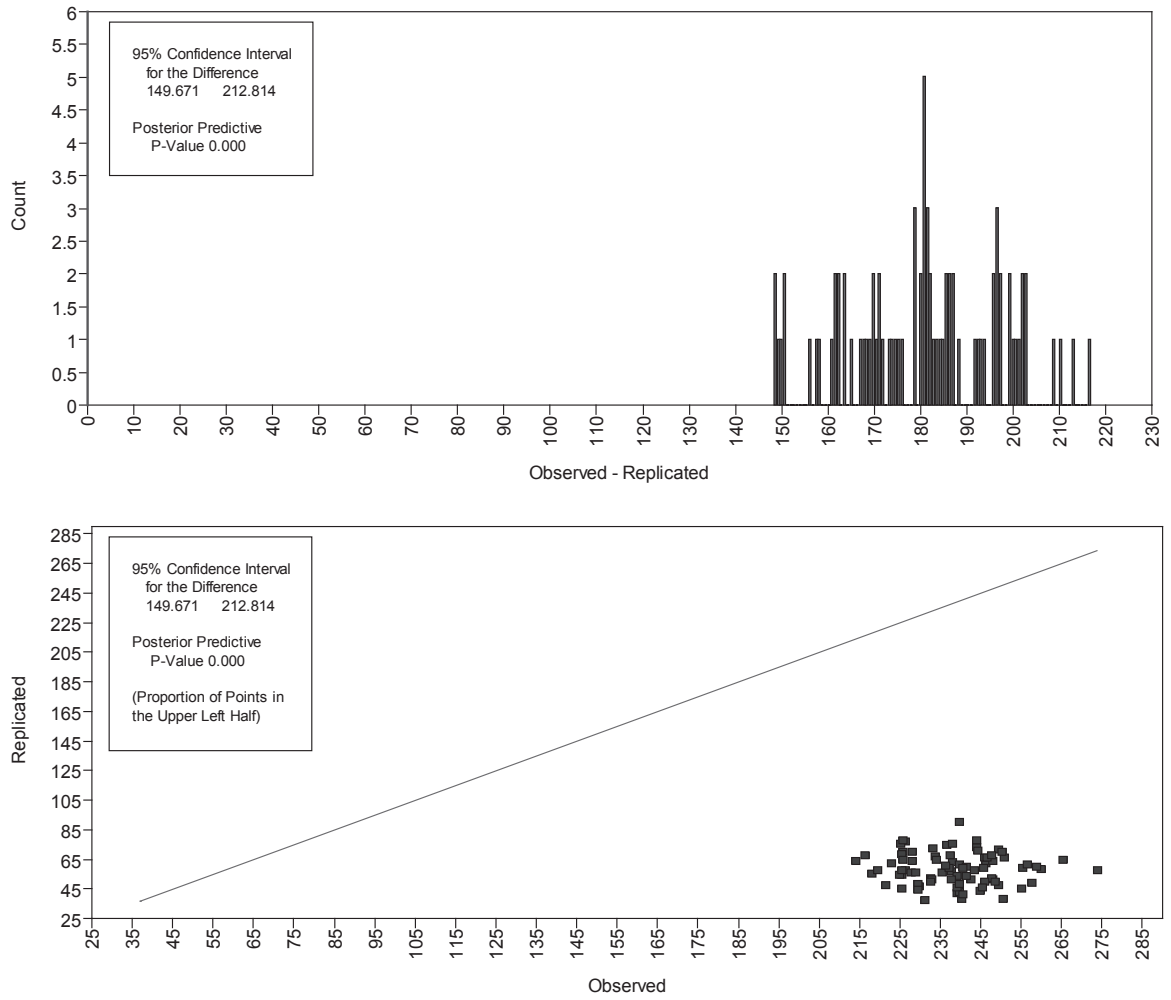

FIGURE 38.2. CFA: PPC 95% confidence interval histogram and PPC scatterplot.

As an illustration of model comparison, the original two-factor model was compared to a one-factor model. The DIC value produced for the original two-factor CFA model was 10,533.37. The DIC value produced for the one-factor CFA model was slightly larger at 10,593.10. This indicates that although the difference in DIC values is relatively small, the two-factor model provides a better representation of the data compared to the one-factor model.

Bayesian Multilevel Path Analysis

This example is based on a reanalysis of a multilevel path analysis described in Kaplan, Kim, and Kim (2009). In their study, a multilevel path analysis was employed to study within- and between-school predictors of mathematics achievement using data from 4,498 students from the Program for International Student Assessment (PISA) 2003 survey (Organization for Economic Co-operation and Development [OECD], 2004). The full

multilevel path analysis is depicted in Figure 38.3. The final outcome variable at the student level was a measure of mathematics achievement (MATHSCOR). Mediating predictors of mathematics achievement consisted of whether students enjoyed mathematics (ENJOY) and whether students felt mathematics was important in life (IMPORTNT). Student exogenous background variables included students' perception of teacher qualities (PERTEACH), as well as both parents' educational levels (MOMEDUC and DADEDUC). At the school level, a model was specified to predict the extent to which students are encouraged to achieve their full potential (ENCOURAG). A measure of teachers' enthusiasm for their work (ENTHUSIA) was viewed as an important mediator variable between background variables and encouragement for students to achieve full potential.

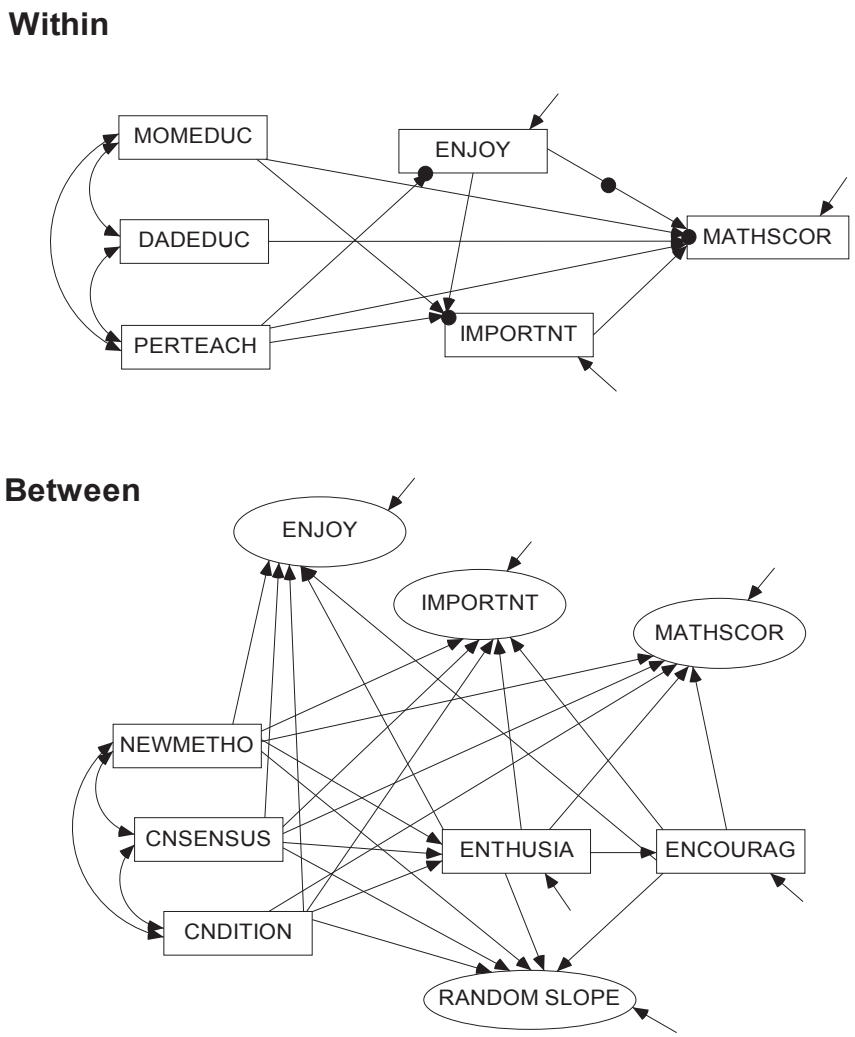

FIGURE 38.3. Multilevel path analysis diagram. Dark circles represent random intercepts and slopes. From Kaplan, Kim, and Kim (2009). Copyright 2009 by SAGE Publications, Inc. Reprinted by permission.

The variables used to predict encouragement via teachers' enthusiasm consisted of math teachers' use of new methodology (NEWMETHO), consensus among math teachers with regard to school expectations and teaching goals as they pertain directly to mathematics instruction (CNSENSUS), and the teaching conditions of the school (CNDITION). The teaching condition variable was computed from the shortage of school's equipment, so higher values on this variable reflect a worse condition.

For this example, we presume to have no prior knowledge of any of the parameters in the model. In this case, all model parameters received normal prior distributions with the mean hyperparameter set at 0 and the variance hyperparameter specified as 10^{10}. The key issue here is the amount of precision in this prior. With this setting, there is very little precision in the prior. As a result, the location of this prior can take on a large number of possible values.

Parameter Convergence

A multilevel path analysis was computed with 5,000 burn-in iterations and 5,000 post-burn-in iterations. The Brooks and Gelman (1998) convergence diagnostic indicated that all parameters properly converged for this model. This model took approximately 1 minute to run.

Figure 38.4 presents convergence plots, posterior density plots, and autocorrelation plots (for both chains) for one of the between-level parameters and one of the within-level parameters. Convergence for these parameters appears to be tight and horizontal, and the posterior probability densities show a close approximation to the normal curve. Finally, the autocorrelation plots are low, indicating that dependence was low for both chains. The additional parameters in this model showed similar results in that convergence plots were tight, density plots were approximately normal, and autocorrelations were low. Appendix 38.2 contains the Mplus code for this example. Note that model fit and model comparison indices are not available for multilevel models and are thus not presented here. This is an area within MCMC estimation that requires further research.

Model Interpretation

Table 38.2 presents selected results for within-level and between-level parameters in the model.[3] For the within-level results, we find that MOMEDUC, DADEDUC, PERTEACH, and IMPORTNT are positive predictors of MATHSCOR. Likewise, ENJOY is positively predicted by PERTEACH. Finally, MOMEDUC, PERTEACH, and ENJOY are positive predictors of IMPORTNT.

The between-level results presented here are for the random slope in the model that relates ENJOY to MATHSCOR. For example, the results indicate that teacher enthusiasm moderates the relationship between enjoyment of mathematics and math achievement, with higher levels of teacher-reported enthusiasm associated with a stronger positive relationship between enjoyment of math and math achievement. Likewise, the math teachers' use of new methodology also demonstrates a moderating effect on the relationship between enjoyment of math and math achievement, where less usage of new methodology lowers the relationship between enjoyment of mathematics and math achievement. The other random slope relationships in the between level can be interpreted in a similar manner.

Bayesian Growth Mixture Modeling

The ECLS-K math assessment data were used for this example (NCES, 2001). Item response theory (IRT) was used to derive scale scores across four time points (assessments were in the fall and spring of kindergarten and first grade) that were used for the growth mixture model. Estimation of growth rates reflects math skill development over the 18 months of the study. The sample for this analysis comprised 592 children and two latent mixture classes.

For this example, we presume to have a moderate degree of prior knowledge of the growth parameters and the mixture class proportions, but no prior knowledge for the factor variances and unique variances. For the growth parameters, we have specified particular location values, but there is only moderate precision defined in the priors (variances = 10). In this case, we are only displaying moderate confidence in the parameter values, as seen through the larger variances specified. This specification provides a wider range of values in the distribution than would be viable but accounts for our lack of strong knowledge through the increased variance term. Stronger knowledge of these parameter values, would decrease the variance hyperparameter term, creating a smaller spread surrounding the location of the prior. However, weaker knowledge of the values would increase the variance term, creating a larger spread surrounding the location of the prior. For the mixture proportions, we presume strong background knowledge

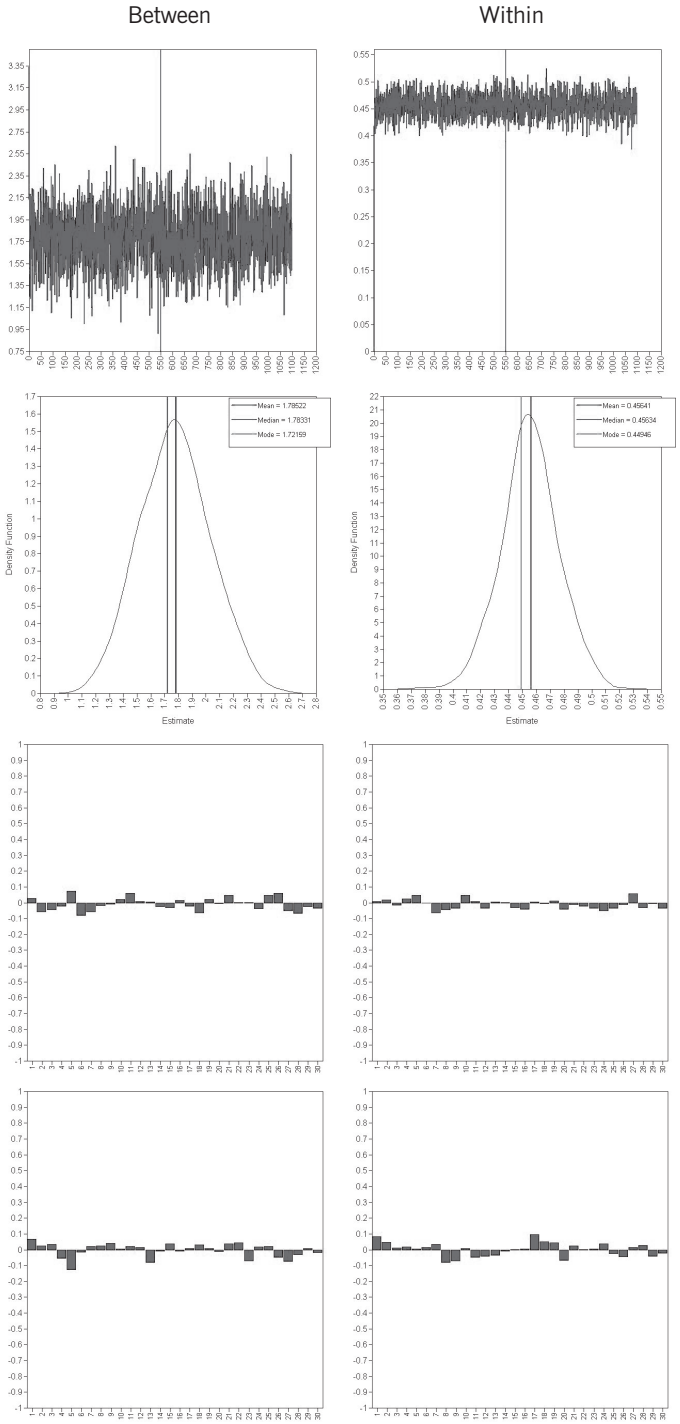

FIGURE 38.4. Multilevel path analysis: Convergence, posterior densities, and autocorrelation plots for select parameters.

TABLE 38.2. Selected MCMC Multilevel Path Analysis Estimates: PISA 2003

Parameter	EAP	*SD*	*p*-value	95% credibility interval
Within level				
MATHSCOR ON MOMEDUC	3.93	0.96	0.00	2.15, 5.79
MATHSCOR ON DADEDUC	4.76	0.96	0.00	2.91, 6.68
MATHSCOR ON PERTEACH	6.10	2.31	0.00	1.64, 10.72
MATHSCOR ON IMPORTNT	15.67	1.98	0.00	11.84, 19.72
ENJOY ON PERTEACH	0.45	0.02	0.00	0.41, 0.49
IMPORTNT ON MOMEDUC	0.02	0.00	0.00	0.01, 0.03
IMPORTNT ON PERTEACH	0.24	0.01	0.00	0.21, 0.27
IMPORTNT ON ENJOY	0.53	0.01	0.00	0.51, 0.55
Between level				
SLOPE ON NEWMETHO	–4.26	2.58	0.05	–9.45, 1.02
SLOPE ON ENTHUSIA	8.95	4.81	0.03	–0.76, 18.23
SLOPE ON CNSENSUS	–3.09	3.72	0.20	–10.65, 4.29
SLOPE ON CNDITION	–8.24	2.66	0.00	–13.53, –3.09
SLOPE ON ENCOURAG	–2.06	2.79	0.23	–7.59, 3.58

Note. EAP, expected a posteriori; *SD*, standard deviation.

of the mixture proportions by specifying class sizes through the Dirichlet prior distribution. The factor variances and unique variances received IW priors that reflected no prior knowledge of the parameter values, as specified in Asparouhov and Muthén (2010).

Parameter Convergence

A growth mixture model was computed, with a total of 10,000 iterations with 5,000 burn-in iterations and 5,000 post-burn-in iterations. The model converged properly, signifying that the Brooks and Gelman (1998) convergence diagnostic indicated parameter convergence for this model. This model took less than 1 minute to run.

Figure 38.5 presents convergence plots, posterior density plots, and autocorrelation plots (for both chains) for the mixture class proportions. Convergence for the mixture class parameters appears to be tight and horizontal. The posterior probability densities show a close approximation to the normal curve. Finally, the autocorrelation plots are quite low, indicating relative sample independence for these parameters for both MCMC chains. The additional parameters in this model showed similar results to the mixture class parameters in that convergence plots were tight, density plots were approximately normal, and autocorrelations were low. Appendix 38.3 contains the Mplus code for this example.

Model Interpretation

The growth mixture model estimates can be found in Table 38.3. For this model, the mean math IRT score for the first latent class (mixture) in the fall of kindergarten was 32.11 and the average rate of change between time points was 14.28. The second latent class consisted of an average math score of 18.75 in the fall of kindergarten, and the average rate of change was 10.22 points between time points. This indicates that Class 1 comprised children with stronger math abilities than Class 2 in the fall of kindergarten. Likewise, Class 1 students also have a larger growth rate between assessments. Overall, 14% of the sample was in the first mixture class, and 86% of the sample was in the second mixture class.

Model Fit

Theory suggests that model comparison via the DIC is not appropriate for mixture models (Celeux, Hurn, & Robert, 2000). As a result, only comparisons from the PPC test will be presented for this growth mixture modeling (GMM) example. Figure 38.6 includes the PPC distribution corresponding to the 95% confidence interval for the difference between the observed data test statistic and the replicated data test statistic. The lower bound of this interval was 718.25, and the upper

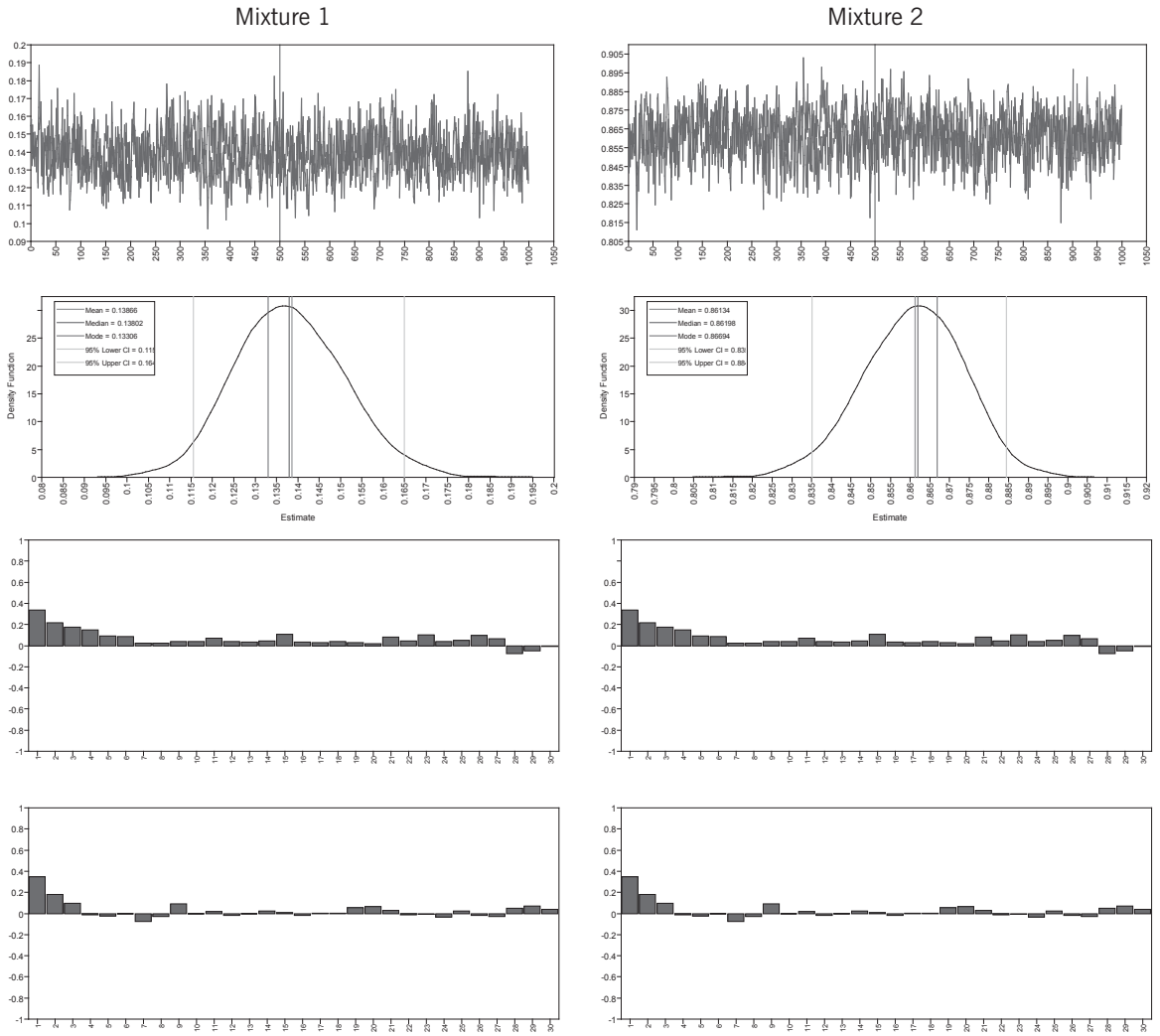

FIGURE 38.5. GMM: Convergence, posterior densities, and autocorrelation plots for mixture class proportions.

bound was 790.56. Similar to the CFA example presented earlier, this positive confidence interval indicates that the observed data test statistic is much larger than what would have been generated by the model. Likewise, Figure 38.6 also includes the PPC scatterplot. All of the points fall below the 45 degree line, which indicates that the model was rejected based on a sufficiently small *p*-value of .00. The results of the PPC test indicate substantial model misfit for this GMM model.

DISCUSSION

This chapter has sought to present an accessible introduction to Bayesian SEM. An overview of Bayesian concepts, as well as a brief introduction to Bayesian computation, was also provided. A general framework of Bayesian computation within the Bayesian SEM framework was also presented, along with three examples covering first- and second-generation SEM.

TABLE 38.3. Mplus MCMC GMM Estimates: ECLS-K Math IRT Scores

Parameter	EAP	*SD*	*p*-value	95% credibility interval
Latent class 1				
Class proportion	0.14			
Intercept and slope correlation	–0.06	0.19	0.38	–0.44, 0.32
Growth parameter means				
Intercept	32.11	1.58	0.00	28.84, 35.09
Slope	14.28	0.78	0.00	12.72, 15.77
Variances				
Intercept	98.27	26.51	0.00	54.37, 158.07
Slope	18.34	4.51	0.00	10.60, 27.76
Latent class 2				
Class proportion	0.86			
Intercept and slope correlation	0.94	0.03	0.00	0.87, 0.98
Growth parameter means				
Intercept	18.75	0.36	0.00	17.98, 19.40
Slope	10.22	0.19	0.00	9.86, 10.61
Variances				
Intercept	22.78	3.63	0.00	16.12, 30.56
Slope	7.84	1.15	0.00	5.93, 10.29
Residual variances				
All time points and classes	32.97	1.17	0.00	30.73, 35.34

Note. EAP, expected a posteriori; *SD*, standard deviation.

With the advent of open-source software for Bayesian computation, such as packages found in R (R Development Core Team, 2008) and WinBUGS (Lunn et al., 2000), as well as the newly available MCMC estimator in Mplus (Muthén & Muthén, 2010), researchers can now implement Bayesian methods for a wide range of research problems.

In our examples, we specified different degrees of prior knowledge for the model parameters. However, it was not our intention in this chapter to compare models under different specification of prior distributions, nor to compare results to conventional frequentist estimation methods. Rather, the purpose of these examples was to illustrate the use and interpretation of Bayesian estimation results.

The relative ease of Bayesian computation in the SEM framework raises the important question of why one would choose to use this method—particularly when it can often provide results that are very close to that of frequentist approaches such as maximum likelihood. In our judgment, the answer lies in the major distinction between the Bayesian approach and the frequentist approach, that is, in the elicitation, specification, and incorporation of prior distributions on the model parameters.

As pointed out by Skrondal and Rabe-Hesketh (2004, p. 206), there are four reasons why one would adopt the use of prior distributions—one of which they indicate is "truly" Bayesian, while the others represent a more "pragmatic" approach to Bayesian inference. The truly Bayesian approach would specify prior distributions that reflect elicited prior knowledge. For example, in the context of SEM applied to educational problems, one might specify a normal prior distribution on the regression coefficient relating socioeconomic status (SES) to achievement, where the hyperparameter on the mean of the regression coefficient is obtained from previous research. Given that an inspection of the literature suggests roughly the same values for the regression coefficient, a researcher might specify a small value for the

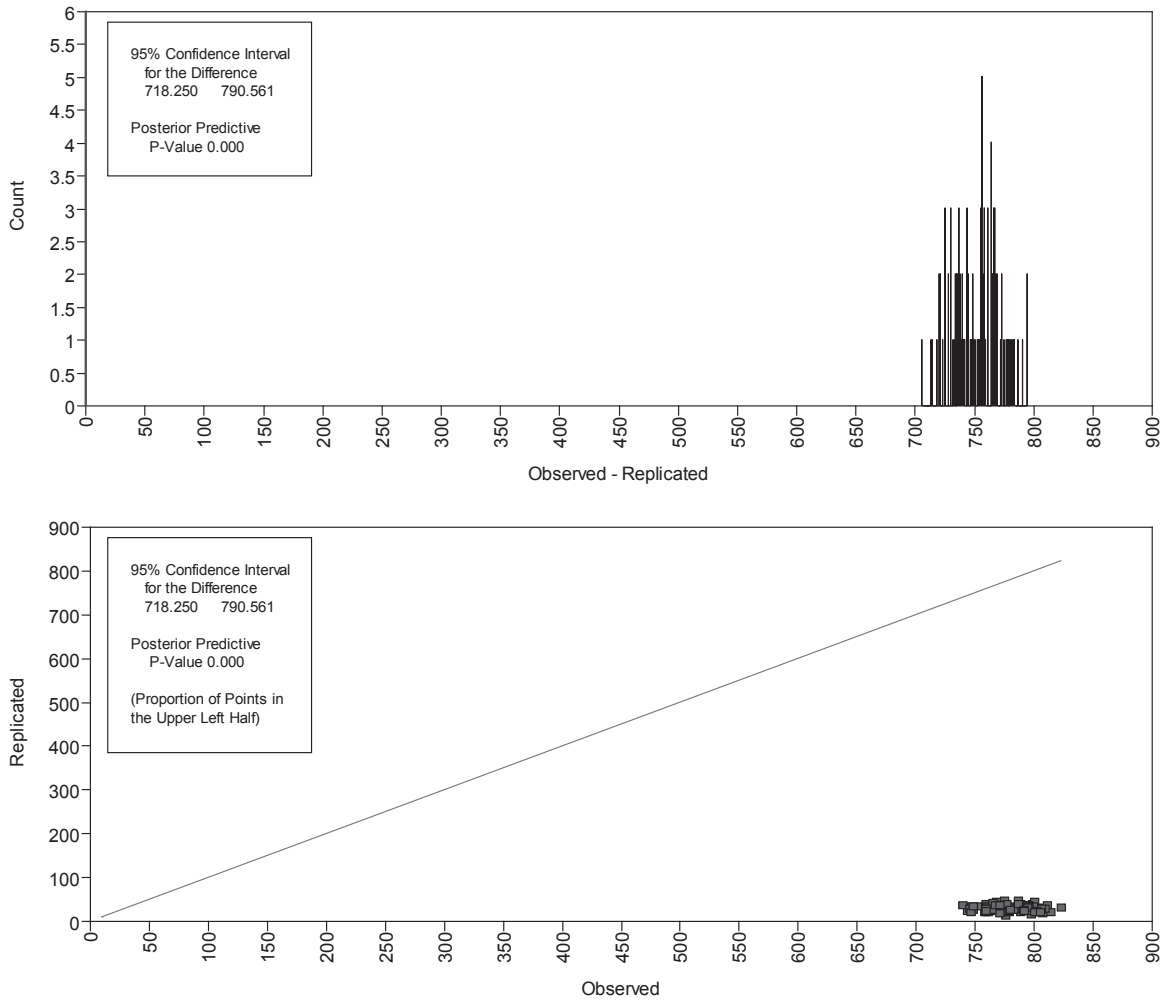

FIGURE 38.6. GMM: PPC 95% confidence interval histogram and PPC scatterplot.

variance of the regression coefficient—reflecting a high degree of precision. Pragmatic approaches, on the other hand, might specify prior distributions for the purposes of achieving model identification, constraining parameters so they do not drift beyond their boundary space (e.g., Heywood cases) or simply because the application of MCMC can sometimes make problems tractable that would otherwise be very difficult in more conventional frequentist settings.

Although we concur with the general point that Skrondal and Rabe-Hesketh (2004) are making, we do not believe that the distinction between "true" Bayesians versus "pragmatic" Bayesians is necessarily the correct distinction to be made. If there is a distinction to be made, we argue that it is between Bayesians and pseudo-Bayesians, where the latter implement MCMC as "just another estimator." Rather, we adopt the pragmatic perspective that the usefulness of a model lies in whether it provides good predictions. The specification of priors based on subjective knowledge can be subjected to quite pragmatic procedures in order to sort out the best predictive model, such as the use of PPC.

What Bayesian theory forces us to recognize is that it is possible to bring in prior information on the distribution of model parameters, but that this requires a deeper understanding of the elicitation problem (see Abbas, Budescu, & Gu, 2010; Abbas, Budescu, Yu, & Haggerty, 2008; O'Hagan et al., 2006). The general idea is that through a careful review of prior research on a problem, and/or the careful elicitation of prior knowledge from experts and/or key stakeholders, relatively precise values for hyperparameters can be obtained and incorporated into a Bayesian specification. Alternative elicitations can be directly compared via Bayesian model selection measures as described earlier. It is through (1) the careful and rigorous elicitation of prior knowledge, (2) the incorporation of that knowledge into our statistical models, and (3) a rigorous approach to the selection among competing models that a pragmatic *and* evolutionary development of knowledge can be realized—and this is precisely the advantage that Bayesian statistics, and Bayesian SEM in particular, has over its frequentist counterparts. Now that the theoretical and computational foundations have been established, the benefits of Bayesian SEM will be realized in terms of how it provides insights into important substantive problems.

ACKNOWLEDGMENTS

The research reported in this chapter was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant No. R305D110001 to the University of Wisconsin–Madison. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

We wish to thank Tihomir Asparouhov and Anne Boomsma for valuable comments on an earlier draft of this chapter.

NOTES

1. The *credibility interval* (also referred to as the *posterior probability interval*) is obtained directly from the quantiles of the posterior distribution of the model parameters. From the quantiles, we can directly obtain the probability that a parameter lies within a particular interval. This is in contrast to the frequentist *confidence interval*, where the interpretation is that $100(1 - \alpha)\%$ of the confidence intervals formed a particular way capture the true parameter of interest under the null hypothesis.
2. Note that in the case where there is only one element in the block, the prior distribution is assumed to be inverse-gamma, that is, $\boldsymbol{\theta}_{IW} \sim IG(a, b)$.
3. Tables with the full results from this analysis are available upon request.

REFERENCES

Abbas, A. E., Budescu, D. V., & Gu, Y. (2010). Assessing joint distributions with isoprobability countours. *Management Science*, *56*, 997–1011.

Abbas, A. E., Budescu, D. V., Yu, H.-T., & Haggerty, R. (2008). A comparison of two probability encoding methods: Fixed probability vs. fixed variable values. *Decision Analysis*, *5*, 190–202.

Albert, J. (2007). *Bayesian computation with R*. New York: Springer.

Asparouhov, T., & Muthén, B. (2010). *Bayesian analysis using Mplus: Technical implementation*. Available from *http://www.statmodel.com/download/Bayes3.pdf*.

Box, G., & Tiao, G. (1973). *Bayesian inference in statistical analysis*. New York: Addison-Wesley.

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

Celeux, G., Hurn, M., & Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association, 95*, 957–970.

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

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis, second edition*. London: Chapman & Hall.

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

Gelman, A., & Rubin, D. B. (1992b). A single series from the Gibbs sampler provides a false sense of security. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 4* (pp. 625–631). Oxford, UK: Oxford University Press.

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

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 4* (pp. 169–193). Oxford, UK: Oxford University Press.

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

Gill, J. (2002). *Bayesian methods*. Boca Raton, FL: CRC Press.

Heidelberger, P., & Welch, P. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, *31*, 1109–1144.

Hoff, P. D. (2009). *A first course in Bayesian statistical methods*. New York: Springer.

Jo, B., & Muthén, B. (2001). Modeling of intervention effects with noncompliance: A latent variable modeling approach for randomized trials. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 57–87). Mahwah, NJ: Erlbaum.

Jöreskog, K. G. (1973). A general method for estimating a linear structural equation system. In A. S. Goldberger & O. D. Duncan (Eds.), *Structural equation models in the social sciences* (pp. 85–112). New York: Academic Press.

Kaplan, D. (2003). Methodological advances in the analysis of individual growth with relevance to education policy. *Peabody Journal of Education*, *77*, 189–215.

Kaplan, D. (2009). *Structural equation modeling: Foundations and extensions* (2nd ed.). Newbury Park, CA: Sage.

Kaplan, D., & Depaoli, S. (in press). Bayesian statistical methods. In T. D. Little (Ed.), *Oxford handbook of quantitative methods*. Oxford, UK: Oxford University Press.

Kaplan, D., Kim, J.-S., & Kim, S.-Y. (2009). Multilevel latent variable modeling: Current research and recent developments. In R. E. Millsap & A. Maydeu-Olivares (Eds.), *The SAGE handbook of quantitative methods in psychology* (pp. 595–612). Newbury Park, CA: Sage.

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

Lee, S.-Y. (1981). A Bayesian approach to confirmatory factor analysis. *Psychometrika*, *46*, 153–160.

Lee, S.-Y. (2007). *Structural equation modeling: A Bayesian approach*. New York: Wiley.

Lunn, D., Thomas, A., Best, N., & Spiegelhalter, D. (2000). Winbugs—a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325–337.

Martin, A. D., Quinn, K. M., & Park, J. H. (2010, May 10). *Markov chain Monte Carlo (MCMC) package*. Available online at *http://mcmcpack.wustl.edu*.

Martin, J. K., & McDonald, R. P. (1975). Bayesian estimation in unrestricted factor analysis: A treatment for Heywood cases. *Psychometrika*, *40*, 505–517.

Muthén, B. (2001). Second-generation structural equation modeling with a combination of categorical and continuous latent variables: New opportunities for latent class/latent growth modeling. In L. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 289–322). Washington, DC: American Psychological Association.

Muthén, B. (2010). *Bayesian analysis in Mplus: A brief introduction*. Available from *http://www.statmodel.com/download/introbayesversion%203.pdf*.

Muthén, B., & Asparouhov, T. (in press). Bayesian SEM: A more flexible representation of substantive theory. *Psychological Methods*.

Muthén, B., & Masyn, K. (2005). Mixture discrete-time survival analysis. *Journal of Educational and Behavioral Statistics*, *30*, 27–58.

Muthén, L. K., & Muthén, B. (2010). *Mplus: Statistical analysis with latent variables*. Los Angeles: Authors.

National Center for Education Statistics (NCES). (2001). *Early childhood longitudinal study: Kindergarten class of 1998–99: Base year public-use data files user's manual* (Tech. Rep. No. NCES 2001-029). Washington, DC: U.S. Government Printing Office.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., et al. (2006). *Uncertain judgements: Eliciting experts' probabilities*. West Sussex, UK: Wiley.

Organization for Economic Cooperation and Development (OECD). (2004). *The PISA 2003 assessment framework: Mathematics, reading, science, and problem solving knowledge and skills*. Paris: Author.

Press, S. J. (2003). *Subjective and objective Bayesian statistics: Principles, models, and applications* (2nd ed.). New York: Wiley.

R Development Core Team. (2008). R: A language and environment for statistical computing [Computer software manual]. Vienna: R Foundation for Statistical Computing. Available from *http://www.R-project.org*.

Raftery, A. E. (1995). Bayesian model selection in social research (with discussion). In P. V. Marsden (Ed.), *Sociological methodology* (Vol. 25, pp. 111–196). New York: Blackwell.

Raftery, A. E., & Lewis, S. M. (1992). How many iterations in the Gibbs sampler? In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 4* (pp. 763–773). Oxford, UK: Oxford University Press.

Scheines, R., Hoijtink, H., & Boomsma, A. (1999). Bayesian estimation and testing of structural equation models. *Psychometrika*, *64*, 37–52.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Sinharay, S. (2004). Experiences with Markov chain Monte Carlo convergence assessment in two psychometric examples. *Journal of Educational and Behavioral Statistics*, *29*, 461–488.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society B*, *64*, 583–639.

APPENDIX 38.1. CFA Mplus Code

```
title: MCMC CFA with ECLS-K math data
data: file is cfadata.dat;
variable: names are y1-y9;
analysis:
    estimator = BAYES; !This option uses the MCMC Gibbs sampler as a default
    chains = 2; !Two chains is the default in Mplus Version 6
    distribution = 10,000; !The first half of the iterations is always used as burn-in
    point = mean; !Estimating the median is the default for Mplus
model priors: !This option allows for priors to be changed from default values
    a2 ~ N(.8,.01); !Normal prior on Factor 1 loading: Item 2
    b4 ~ N(.8,.01); !Normal prior on Factor 2 loading: Item 4
    b5 ~ N(.8,.01); !Normal prior on Factor 2 loading: Item 5
    b6 ~ N(.8,.01); !Normal prior on Factor 2 loading: Item 6
    b7 ~ N(.8,.01); !Normal prior on Factor 2 loading: Item 7
    b8 ~ N(.8,.01); !Normal prior on Factor 2 loading: Item 8
    b9 ~ N(.8,.01); !Normal prior on Factor 2 loading: Item 9
model:
    f1 by y1@1 y2*.8(a2); !Normal priors on Factor 1 loadings with arbitrary item identifiers (a2)
    f2 by y3@1 y4-y9*.8(b4-b9); !Priors on Factor 2 loadings with arbitrary item identifiers (b4-b9)
    f1*1;
    f2*1;
    f1 with f2 *.4;
plot:
    type = plot2; !Requesting all MCMC plots: convergence, posterior densities, and autocorrelations
```

APPENDIX 38.2. Multilevel Path Analysis with a Varying-Slope Mplus Code

```
title: Path Analysis
data: File is multi-level.dat;
variable: names are schoolid newmetho enthusia cnsensus
    cndition encourag momeduc dadeduc
    perteach enjoy importnt mathscor;
    Usevariables are newmetho enthusia cnsensus
    cndition encourag momeduc dadeduc
    perteach enjoy importnt mathscor;
    Between = newmetho enthusia cnsensus cndition encourag;
    Cluster is schoolid;
analysis: type = twolevel random;
    estimator = BAYES;
    point=mean;
model:
%Within%
    mathscor ON momeduc dadeduc perteach importnt;
    enjoy ON perteach;
    importnt ON momeduc perteach enjoy;
    momeduc WITH dadeduc perteach;
    dadeduc WITH perteach;
    slope | mathscor ON enjoy;
```

(cont.)

APPENDIX 38.2. *(cont.)*

```
%Between%
      mathscor ON newmetho enthusia cnsensus cndition encourag;
      enjoy ON newmetho enthusia cnsensus cndition encourag; importnt ON
newmetho enthusia cnsensus cndition encourag;
      slope ON newmetho enthusia cnsensus cndition encourag;
      encourag ON enthusia;
      enthusia ON newmetho cnsensus cndition;
plot: type=plot2;
```

APPENDIX 38.3. Growth Mixture Model Mplus Code

```
title: MCMC GMM with ECLS-K math data
data: file is Math GMM.dat;
variable: names are y1-y4;
      classes =c(2);
analysis:
      type = mixture;
      estimator = BAYES; !This option uses the MCMC Gibbs sampler as a default
      chains = 2; !Two chains is the default in Mplus Version 6
      distribution = 10,000; !The first half of the iterations is always used as burn-in
      point = mean; !Estimating the median is the default for Mplus
model priors: !This option allows for priors to be changed from default values
      a ~ N(28,10); !Normal prior on mixture class 1 intercept
      b ~ N(13,10); !Normal prior on mixture class 1 slope
      c ~ N(17,10); !Normal prior on mixture class 2 intercept
      d ~ N(9,10); !Normal prior on mixture class 2 slope
      e ~ D(80,510); !Dirichlet prior on mixture class proportions
model:
%overall%
      y1-y4*.5;
      i s | y1@0 y2@1 y3@2 y4@3;
      i*1; s*.2;
      [c#1*-1](e); !Setting up Dirichlet prior on mixture class proportions with arbitrary identifier (e)
      y1 y2 y3 y4 (1);
%c#1%
      [i*28](a); !Setting up Normal prior on mixture class 1 intercept with arbitrary identifier (a)
      [s*13](b); !Setting up Normal prior on mixture class 1 slope with arbitrary identifier (b)
      i with s;
      i; s;
%c#2%
      [i*17](c); !Setting up Normal prior on mixture class 2 intercept with arbitrary identifier (c)
      [s*9](d); !Setting up Normal prior on mixture class 2 intercept with arbitrary identifier (d)
      i with s;
      i; s;
plot:
      type = plot2; !Requesting all MCMC plots: convergence, posterior densities, and autocorrelations
      output: stand;
      cinterval;
```


Guilford Publications
72 Spring Street
New York, NY 10012
212-431-9800
800-365-7006
www.guilford.com