23

Analyzing International Large-Scale Assessment Data within a Bayesian Framework

David Kaplan
University of Wisconsin-Madison

Soojin Park
University of Wisconsin-Madison

CONTENTS

Analyzing International Large-Scale Assessment Data within a Bayesian Framework

> . . .it is clear that it is not possible to think about learning from experience and acting on it without coming to terms with Bayes' theorem.
>
> —**Jerome Cornfield**

When listening to the news or reading newspapers, it not uncommon to find periodic surges of interest regarding which countries are performing the best in major academic subject domains such as reading, mathematics, and science. Information on the comparability of educational outcomes across countries is provided by international large-scale assessments (ILSAs) such as the Organization for Economic Cooperation and Development (OECD)-sponsored Program for International Student Assessment (PISA; OECD 2012), the International Association for the Evaluation of Educational Achievement (IEA)-sponsored Trends in International Mathematics and Science Study (TIMSS; Martin et al. 2008), and the IEA-sponsored Progress in International Reading Literacy Study (PIRLS; Mullis et al. 2008).

The policy consequences of these cross-country comparisons can be quite profound. For example, according to the results of PISA 2000, the average score of German students was below the OECD average, and it was found to be lower than other countries that had similar levels of per capita gross domestic product (GDP). Not only did Germany exhibit relatively low performance in reading, but PISA 2000 results also showed inequities in schooling outcomes in terms of the socioeconomic background of students. Subsequent PISA surveys pointed out that the early tracking policies of the German educational system were not associated with better overall performance, but in fact were associated with low equity (OECD 2003b, 2010). The results of PISA 2000 triggered a unilateral decision by the German federal government to agree to educational reforms and national standards, which moved the German curricula toward a more practical focus (Wiseman 2010). Of course, the policy changes enacted by Germany on the basis of PISA 2000 are not guaranteed to produce the same positive results in another country or economy.

In addition to the use of ILSAs for policy analysis, a great deal of basic research has been conducted that utilizes a variety of ILSAs covering different domains of interest as well as different age or grade levels. A review of the extant literature shows that the vast majority of research conducted with PISA, TIMSS, and PIRLS utilizes a range of statistical methods, including simple regression analysis, logistic regression modeling, multilevel modeling, factor analysis, item response theory, and structural equation modeling. In almost every instance, these methods have been conducted within the Fisherian and Neyman/Pearson schools of statistics. These schools constitute the "classical" school of statistics that rest on the foundations of the frequentist view of probability.

In contrast to the frequentist school of statistics, the Bayesian school presents a coherent, internally consistent, and (arguably) more powerful alternative to the classical school. However, Bayesian statistics has long been ignored in the quantitative methods training of social scientists. Typically, the only introduction that a student might have to Bayesian methodology is a brief overview of Bayes' theorem while studying probability in an introductory statistics class. Two reasons can be given for lack exposure to Bayesian methods. First, until recently, it was not possible to conduct statistical modeling from a Bayesian perspective because of its complexity and lack of available software. Second, Bayesian statistics represents a compelling alternative to frequentist statistics, and is, therefore, controversial. However, in recent years, there has been renewed interest in the Bayesian alternative along with extraordinary developments in the extension and application of Bayesian statistical methods to the social and behavioral sciences. This growth has been attributed mostly to developments in powerful computational tools that render the specification and estimation of complex models such as those used in the analysis of ILSA data feasible from a Bayesian perspective. However, and in addition, growing interest in Bayesian inference has also stemmed from an overall dissatisfaction with the internal inconsistencies of the classical approach (Howson and Urbach 2006).

The orientation of this chapter will be toward those who conduct research using ILSA data and who are well trained in statistical modeling within the frequentist paradigm. The goal is to introduce concepts of model specification, estimation, and evaluation from the Bayesian perspective. Nevertheless, the scope of this chapter is, by necessity, limited, because the field of Bayesian statistics is remarkably wide ranging and space limitations preclude a full development of Bayesian theory. Moreover, not all concepts covered in this chapter will be demonstrated in our examples, owing to present software limitations. Thus, the goal of this chapter will be to lay out the fundamental ideas of Bayesian statistics as they pertain specifically to the analysis if ILSA data.

The organization of this chapter is as follows. The first section provides an overview of Bayesian statistical inference. This is followed by a section describing how Bayesian models are evaluated and compared. Next, we briefly discuss the elements of Bayesian computation. We then present three examples of common analyses conducted on ILSA data from a Bayesian perspective. Our examples will utilize the PISA study, but the issues and implications outlined are applicable to TIMSS, PIRLS, and other ILSA endeavors. We will specify models using data from PISA 2009, comparing the case where priors are not used to the case where priors based on results from PISA 2000 are used. First, we will specify a simple country-level regression to examine the influence of priors in a small sample size case. Next, we will estimate a hierarchical linear model using U.S. PISA data. Finally, we will estimate a school-level confirmatory factor analysis model of school administrator perceptions of school climate, again using U.S. PISA data. All analyses utilize the Mplus software program (Muthén and Muthén 2012).

Overview of Bayesian Statistical Inference

This section provides an overview of Bayesian inference and follows closely the recent overview by Kaplan and Depaoli (2012a) here discussed within the context of ILSA data generally, and, in particular, PISA.

To begin, denote by Y a random variable that takes on a realized value y. In the context of PISA, Y could be the PISA index of economic, social, and cultural status (ESCS).* In the context of more advanced methods, Y could be vector-valued, such as items on the PISA school climate survey. Once the student 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, we 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 the ESCS distribution, or it can be vector valued, such as the set of all parameters of a factor analysis of the PISA school climate survey.

We are concerned with determining the probability of observing y given the unknown parameters θ, which we write as $p(y \mid \theta)$. 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 $L(\theta \mid y)$. Often, we work with the log-likelihood written as $l(\theta \mid y)$. In accordance with the likelihood principle (see, e.g., Royall 1997), the likelihood function summarizes all of the statistical information in the data.

The key difference between Bayesian statistical inference and frequentist statistical inference concerns the nature of the unknown parameters θ. In the frequentist framework, θ is assumed to be unknown, but fixed. In Bayesian statistical inference, any quantity that is unknown, such as θ, is considered to be 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 considered to be random, we can model the joint probability of the parameters and the data as a function of the conditional probability distribution of the data given the parameters, and the prior probability distribution of the parameters. More formally

$$p(\theta, y) = p(y|\theta)p(\theta). \tag{23.1}$$

Because of the symmetry of joint probabilities

$$p(y|\theta)p(\theta) = p(\theta|y)p(y). \tag{23.2}$$

* From the OECD glossary of statistical terms, ESCS "was created on the basis of the following variables: the International Socio-Economic Index of Occupational Status (ISEI); the highest level of education of the students parents, converted into years of schooling; the PISA index of family wealth; the PISA index of home educational resources; and the PISA index of possessions related to 'classical' culture in the family home."

Therefore

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(y|\theta)p(\theta)}{p(y)} \tag{23.3}$$

where $p(\theta|y)$ is referred to as the *posterior distribution* of the parameters θ given the observed data y. Thus, from Equation 23.3, 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 23.3 is *Bayes' theorem*. For discrete variables

$$p(y) = \sum_{\theta} p(y|\theta)p(\theta), \tag{23.4}$$

and for continuous variables

$$p(y) = \int_{\theta} p(y|\theta)p(\theta)\, d\theta. \tag{23.5}$$

As above, the denominator in Equation 23.3 does not involve model parameters so we can omit the term and obtain the *unnormalized posterior distribution*

$$p(\theta \mid y) \propto p(y \mid \theta)p(\theta) \tag{23.6}$$

or equivalently

$$p(\theta \mid y) \propto L(\theta \mid y)p(\theta). \tag{23.7}$$

Equation 23.6 represents the core of Bayesian statistical inference and it is what separates Bayesian statistics from frequentist statistics. Specifically, Equation 23.6 states that our uncertainty regarding the parameters of our model, as expressed by the prior distribution $p(\theta)$, is *moderated* by the actual data $p(y \mid \theta)$, or equivalently, $L(\theta|y)$, yielding an updated estimate of the model parameters as expressed by the posterior distribution $p(\theta|y)$. Again, in the context of the ESCS index, Equation 23.6 states that the posterior distribution of the parameters underlying ESCS (e.g., the mean and/or the variance) is proportional to the prior information based, perhaps, on previous research, moderated by the actual sample data on ESCS as summarized by the likelihood function. The Bayesian framework thus encodes our prior knowledge of the parameters via the prior distribution. Updated knowledge

of the parameters of our model is obtained from a summary of the posterior distribution. Summaries of the parameters of the posterior distribution (described later) can (and should) be used as new priors in a subsequent study. In essence, this is how the Bayesian framework supports evolutionary knowledge development and it is what separates it from the frequentist school of statistics.

Important Assumption: Exchangeability

In most discussions of likelihood, and indeed, in the specification of most statistical models, it is common to invoke the assumption that the data $y_1,y_2,...,y_n$ are independently and identically distributed—often referred to as the *i.i.d.* assumption. Bayesians, however, invoke the deeper notion of *exchangeability* to produce likelihoods and address the issue of independence.

Exchangeability arises from de Finetti's representation theorem (de Finetti, 1974) and implies that the subscripts of a vector of data, for example, $y_i,y_2,...,y_n$, do not carry information that is relevant to describing the probability distribution of the data. In other words, the joint distribution of the data, $f(y_i,y_2,...,y_n)$ is invariant to permutations of the subscripts.*

As a simple example of exchangeability, consider the response that student i would have to the question appearing in PISA 2009, "How much do you agree or disagree with each of the following statements about teachers at your school?"—"Most teachers are interested in my well-being," where for simplicity we recode the responses as

$$yi = \begin{cases} 1, & \text{if student } i \text{ agrees} \\ 0, & \text{if student } i \text{ disagrees.} \end{cases} \tag{23.8}$$

Next, consider three possible responses to 10 randomly selected students

$$p(1,0,1,1,0,1,0,1,0,0)$$
$$p(1,1,0,0,1,1,1,0,0,0)$$
$$p(1,0,0,0,0,0,1,1,1,1)$$

Exchangeability implies that only the total number of agreements matter, not the location of those agreements in the vector. This is a subtle assumption insofar as it means that we believe that there is a parameter θ that generates the observed data via a stochastic model and that we can describe that parameter without reference to the particular data at hand (Jackman 2009).

* Technically, according to de Finetti (1974), this refers to *finite* exchangeability. Infinite exchangeability is obtained by adding the provision that every finite subset of an infinite sequence is exchangeable.

As Jackman (2009) points out, the fact that we can describe θ without reference to a particular set of data is, in fact, what is implied by the idea of a prior distribution. In fact, as Jackman notes, "the existence of a prior distribution over a parameter is a *result* of de Finetti's Representation Theorem (de Finetti 1974), rather than an assumption" (Jackman 2009, p. 40).

It is important to note that exchangeability is weaker than the statistical assumption of independence. In the case of two events—say A and B—independence implies that $p(A \mid B) = p(A)$. If these two events are independent, then they are exchangeable—however, exchangeability does not imply independence.

A generalization of de Finetti's representation theorem that is of relevance to the analysis of ILSA data relates to the problem of *conditional exchangeability*. In considering the PISA 2009 example, a more realistic situation arises because students are nested in schools. Thus, exchangeability at the student level would not be expected to hold, because in considering the entire sequence of responses, school subscripts ($g = 1,2,...,G$) on the individual response (e.g., y_{ig}) are not exchangeable. However, within a given school, students might be exchangeable, and the schools themselves (within a country) might be exchangeable. This issue also leads to the more general idea of Bayesian hierarchical models, in which case exchangeability applies not just to data but also to parameters (Jackman 2009). In our multilevel modeling example below, we will assume at least conditional exchangeability.

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

The argument being made throughout this chapter is that prior cycles of PISA, and indeed information gleaned from other cognate surveys, can provide a knowledge base to be used in subsequent studies. However, in some cases, a researcher may lack, or be unwilling to specify, 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 models. In other words, it is equally 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 distribution into our specification. Noninformative

prior distributions are also referred to as *objective*, *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 [−∞,∞] distribution would be an *improper* prior distribution because 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), and a discussion of "objective" Bayesian inference can be found in Berger (2006).

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. A conjugate prior distribution is one that, when combined with the likelihood function, yields a posterior 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 not have a form that is analytically simple to solve. Arguably, the existence of numerical simulation methods for Bayesian inference, such as Markov chain Monte Carlo 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 relatively straightforward to fully describe its components—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

$$E(\theta|y) = \int_{-\infty}^{+\infty} \theta p(\theta|y)\, d\theta, \tag{23.9}$$

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)

$$\begin{aligned} \text{var}(\theta|y) &= E\left[(\theta - E\left[\left(\theta|y\right]\right)^2|y\right), \\ &= E(\theta^2|y) - E\left(\theta|y\right)^2. \end{aligned} \tag{23.10}$$

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; the so-called *maximum a posteriori* (MAP) estimate. Those measures, along with the quantiles of the posterior distribution, provide a complete description of the distribution.

Posterior Probability 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 $\bar{y}$ 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 *posterior probability interval* (PPI; also known as a *credible interval*). Again, because we assume that an unknown parameter can be described by 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.

Formally, a 100(1 − α)% PPI for a particular subset of the parameter space θ is defined as

$$1 - \alpha = \int_C p\left(\theta|y\right) d\theta. \tag{23.11}$$

So, for example, a 95% PPI means that the probability that the parameter lies in the interval is 0.95. Notice that the interpretation of the PPI is entirely different from the frequentist confidence interval.

Bayesian Model Evaluation and Comparison

Posterior Predictive Model Checking

An important aspect of Bayesian model evaluation that sets it apart from its frequentist counterpart is its focus on posterior prediction. The general idea behind posterior predictive model 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 model 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 model checking utilizes the posterior predictive distribution of replicated data. Following Gelman et al. (2003), let y^{rep} be data replicated from our current model. That is

$$p(y^{rep}|y) = \int p(y^{rep}|\theta)p(\theta|y)\,d\theta$$
$$= \int p(y^{rep}|\theta)p(y|\theta)p(\theta)\,d\theta. \quad (23.12)$$

Notice that the second term, $p(\theta|y)$, on the right-hand side of Equation 23.12 is simply the posterior distribution of the model parameters. In the context of PISA, Equation 23.12 states that given current data y, the distribution of future observations on, say, a model predicting student reading scores from background characteristics, denoted as $p(y^{rep}|y)$, is equal to the probability distribution of the future observations based on the model given the parameters, $p(y^{rep}|\theta)$, weighted by the posterior distribution of the model parameters, $p(y|\theta)p(\theta)$. Thus, posterior predictive checking accounts for uncertainty in both the parameters underlying the model and uncertainty in the data itself.

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 statistic that can be used to measure the discrepancy between the observed data and replicated data is the likelihood ratio chi-square statistic. In Mplus (Muthén and Muthén 2012), each draw of the posterior estimates is used to generate replicated data. Then, the likelihood ratio chi-square is computed comparing the observed data to the replicated data for each draw. A scatterplot can be drawn that displays the likelihood ratio for the replicated data against the likelihood ratio for the observed data.

An approach to summarizing posterior predictive checking incorporates the notion of Bayesian p-values. Denote by $T(y)$ the likelihood ratio test statistic

based on the data and the model parameters estimated at the tth MCMC iteration. Further, let $T(y^{rep})$ be the same test statistic but defined for the replicated data based on the model parameters estimated at the tth MCMC iteration (described below). Then, the Bayesian p-value is defined to be

$$p\text{-value} = p(T((y) < T(y^{rep})y). \quad (23.13)$$

Lower values of Equation 23.13 suggest poor model fit insofar as the test statistic based on the data does not equal or exceed the test statistic based on replicated data generated from the model parameters themselves. Mplus will produce a 95% confidence interval around the difference between $T(y)$ and $T(y^{rep})$. If the lower limit of the confidence interval is positive, it suggests poor model fit. Good model fit is considered to be a Bayesian p-value of approximately 0.50 (Muthén and Asparouhov, 2012a). Mplus will also produce a posterior predictive checking scatterplot, where the number of points above the 45-degree line corresponds to the Bayesian p-value. We will demonstrate posterior predictive checking in our examples.

Deviance Information Criterion

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 chosen model will be used for prediction. For this chapter, we will focus on the *deviance information criterion* (DIC; Spiegelhalter et al. 2002) as a method for choosing among a set of competing models.

The DIC is one of many different types of *information criteria* for model selection. Arguably, the most popular method for model selection is the *Bayesian information criterion*. The BIC is derived from so-called *Bayes factors* (Kass and Raftery 1995). In essence, a Bayes factor provides a way to quantify the odds that the data favor one hypothesis over another, where the hypotheses do not need to be nested. When the prior odds of favoring one hypothesis over another are equal, the Bayes factor reduces to the ratio of two integrated likelihoods. The BIC can then be derived from this ratio (see Kass and Raftery 1995; Raftery 1995).

Although the BIC is derived from a fundamentally Bayesian perspective, it is often productively used for model comparison in the frequentist domain. However, the DIC is an explicitly Bayesian approach to model comparison that was developed based on the notion of *Bayesian deviance*. Consider a particular model proposed for a set of data, denoted as $p(y|\theta)$. Then, we begin by defining *Bayesian deviance* as

$$D(\theta) = -2\log\left[p(y|\theta)\right] + 2\log\left[h(y)\right]. \quad (23.14)$$

where 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 23.14 is similar to the BIC, it is not, as currently defined, an explicitly Bayesian measure of model fit. To accomplish this, we use Equation 23.14 to obtain a posterior mean over θ by defining

$$\overline{D(\theta)} = E_\theta\left[-2\log\left[p(y \mid \theta) \mid y\right] + 2\log\left[h(y)\right]\right]. \tag{23.15}$$

Next, let $D(\bar{\theta})$ be a posterior estimate of θ. From here, we can define the *effective dimension* of the model as

$$q_D = \overline{D(\theta)} - D(\bar{\theta}), \tag{23.16}$$

which is the mean deviance minus the deviance of the means. Notice that q_D is a Bayesian measure of model complexity. With q_D in hand, we simply add the model fit term $\overline{D(\theta)}$ to obtain the DIC—namely

$$\text{DIC} = \overline{D(\theta)} + q_D, \tag{23.17}$$

$$= 2\overline{D(\theta)} - D(\bar{\theta}) \tag{23.18}$$

Similar to the BIC, the model with the smallest DIC among a set of competing models is preferred. The DIC is available in Mplus (Muthén and Muthén 2012) and will be demonstrated in the examples below.

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 as well as open-source software. The most common algorithm for Bayesian estimation is based on MCMC sampling. In the interest of space, we will not discuss the details of MCMC sampling and instead refer the reader to number of very important papers and books that have been written about MCMC sampling (see, e.g., Gilks et al. 1996). Suffice 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.

For the purposes of this chapter, we will use the Gibbs sampler (Geman and Geman 1984) as implemented in Mplus (Muthén and Muthén 2012). Informally, the Gibbs sampler proceeds as follows. Consider that the goal is to obtain the joint posterior distribution of two model parameters—say, θ_1 and θ_2, given some data y, written as $f(\theta_1,\theta_2|y)$. These two model parameters can be regression coefficients from a simple multiple regression model. Dropping the conditioning on y for simplicity, what is required is to sample from $f(\theta_1|\theta_2)$ and $f(\theta_2|\theta_1)$. In the first step, an arbitrary value for θ_2 is chosen, say θ_2^0. We next obtain a sample from $f(\theta_1|\theta_2^0)$. Denote this value as θ_1^1. With this new value, we then obtain a sample θ_2^1, from $f(\theta_2|\theta_1^1)$. The Gibbs algorithm continues to draw samples using previously obtained values until two long chains of values for both θ_1 and θ_2 are formed. It is common that the first m of the total set of samples is dropped. These are referred to as the *burn-in* samples. The remaining samples are then considered to be draws from the marginal posterior distributions of $f(\theta_1)$ and $f(\theta_2)$.

An important part of MCMC estimation is assessing the convergence of the algorithm. Here too, a number of approaches exist to determine if the algorithm has converged (see, e.g., Sinharay 2004). A variety of these diagnostics are reviewed and demonstrated in Kaplan and Depaoli (2012a), including the Geweke convergence diagnostic (Geweke 1992), the Heidelberger and Welch convergence diagnostic (Heidelberger and Welch 1983), the Raftery and Lewis convergence diagnostic (Raftery and Lewis 1992), and the Brooks, Gelman, and Rubin diagnostic (Gelman and Rubin 1992a,b; Gelman 1996).

Visual diagnostics of convergence include the trace plot and the autocorrelation plot. The trace plot shows the value of the estimate at the tth iteration of the algorithm. Convergence is indicated by a tight "caterpillar-like" band centered around the modal value of the estimate. The autocorrelation plot shows the degree to which the current value of the parameter is dependent on the immediate value of the parameter. High autocorrelation suggests poor convergence and that the MCMC algorithm did not do a good job of exploring the posterior distribution (see Kim and Bolt 2007). We will present both plots in the examples below.

When implementing the Gibbs sampler with multiple chains, one of the most common diagnostics is the Brooks, Gelman, and Rubin diagnostic (see, e.g., Gelman and Rubin 1992a,b; Gelman 1996). 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 is 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 should 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, which does not include the sampling variability correction. The changes by Brooks and Gelman reflect the diagnostic as implemented in Mplus (Muthén and Muthén 2012). Once it has been determined that the algorithm has converged, summary statistics, including the posterior mean, mode, standard deviation, and PPI, can be obtained.

EXAMPLE 23.1: BAYESIAN REGRESSION ANALYSIS

In this section, we provide an example of Bayesian regression analysis applied to a country-level analysis of reading performance using data from PISA 2000 and 2009. Recall that the year 2000 was the first cycle of PISA and the major domain was reading. PISA 2009 represented the first complete domain cycle of PISA concentrating again on reading. The goal in presenting this example is twofold. First, we wish to demonstrate Bayesian extensions of a commonly used method applied to a sensible question of policy and research relevance. Second, we wish to compare the results of analyses using PISA 2009 when we have no prior information (the noninformative prior case) to the case where we use information gleaned from PISA 2000 to provide informative priors (the informative prior case).

We begin by discussing the basic model with noninformative and informative priors. We then turn to the results, which provide a comparison of choice of priors in the context of a relatively small sample size problem.

Model

Consider a very simple model regressing country-level reading proficiency on country-level background predictors. To begin, let $\mathbf{y}$ be an n-dimensional vector $(y_1,y_2,...,y_n)'$ $(i = 1,2,...,n)$ of scores from n countries on the PISA reading assessment, and let $\mathbf{X}$ be an $n \times k$ matrix containing k background measures, such as GDP, country average teacher salaries, and so on. Then, the normal linear regression model can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \tag{23.19}$$

where $\boldsymbol{\beta}$ is a $k \times 1$ vector of regression coefficients and where the first column of $\boldsymbol{\beta}$ contains an n-dimensional unit vector to capture the intercept term. We assume that country-level PISA reading scores are generated from a normal distribution—specifically

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}), \tag{23.20}$$

where $\mathbf{I}$ is an identity matrix. Moreover, we assume that the n-dimensional vector $\mathbf{u}$ of disturbance terms is assumed to be independently, identically, and normally distributed—specifically

$$\mathbf{u} \sim N\left(0, \sigma^2\mathbf{I}\right). \tag{23.21}$$

From standard linear regression theory, the likelihood of the model parameters $\boldsymbol{\beta}$ and σ^2 can be written as

$$L\left(\boldsymbol{\beta}, \sigma^2 | \mathbf{X}, \mathbf{y}\right) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}. \tag{23.22}$$

Noninformative Priors

In the context of the normal linear regression model, the uniform distribution is typically used as a noninformative prior. That is, we assign an improper uniform prior to the regression coefficient β that allows β to take on values over the support $[-\infty, \infty]$.* This can be written as $p(\beta) \propto c$, where c is a constant.

Next, we assign a uniform prior to $\log(\sigma^2)$ because this transformation also allows values over the support $[0, \infty]$. From here, the joint posterior distribution of the model parameters is obtained by multiplying the prior distributions of $\boldsymbol{\beta}$ and σ^2 by the likelihood give in Equation 23.22. Assuming that $\boldsymbol{\beta}$ and σ^2 are independent, we obtain

$$\begin{aligned} p\left(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, \mathbf{X}\right) &\propto L\left(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, \mathbf{X}\right) p(\boldsymbol{\beta}) p\left(\sigma^2\right), \\ &\propto \left(\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\} \times c \times \sigma^{-2}. \end{aligned}$$

Noting that c does not contain model parameters, and so drops out with the proportionality, we obtain

$$p\left(\boldsymbol{\beta}, \hat{\sigma}^2 | \mathbf{y}, \mathbf{X}\right) \propto \left(\sigma^2\right)^{(-n/2+1)} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\} \tag{23.23}$$

As pointed out by Lynch (2007), the posterior distribution of the model parameters in Equation 23.23 differs from the likelihood only in the leading exponent $(n/2 + 1)$, which although is of no consequence in large samples, may be of consequence in the subsequent example. Here, we see how the Bayesian approach and the frequentist

* The *support* of a distribution refers to the smallest closed interval (or "set" if the distribution is multivariate) whose elements are actually members of the distribution. The complement to this set has elements with probabilities of zero.

approach align when samples are large and priors are noninformative. When samples are small, however, priors can dominate the likelihood and have much greater influence on summaries of the posterior distribution.

Informative Conjugate Priors

Turning to conjugate priors, the most sensible conjugate prior distribution for the vector of regression coefficients $\boldsymbol{\beta}$ of the linear regression model is the multivariate normal prior. The argument for using the multivariate normal distribution as the prior for $\boldsymbol{\beta}$ lies in the fact that the asymptotic distribution of the regression coefficients is normal (Fox 2008).

The conditional prior distribution of the vector $\boldsymbol{\beta}$ given σ^2 can be written as

$$p\left(\boldsymbol{\beta}|\sigma^2\right)=(2\pi)^{k/2}|\Sigma|^{1/2}\exp\left[-\frac{1}{2}(\boldsymbol{\beta}-\mathbf{B})'\Sigma^{-1}(\boldsymbol{\beta}-\mathbf{B})\right], \quad (23.24)$$

where k is the number of variables, $\mathbf{B}$ is the vector of mean hyperparameters assigned to $\boldsymbol{\beta}$, and $\Sigma=\sigma^2\mathbf{I}$ is the diagonal matrix of constant disturbance variances.

The conjugate prior for the variance of the disturbance term σ^2 is the inverse-gamma distribution, with hyperparameters a and b. We write the conjugate prior distribution for σ^2 as

$$p\left(\sigma^2\right)\propto\left(\sigma^2\right)^{-(a+1)}e^{-b/\sigma^2} \quad (23.25)$$

With the likelihood $L(\boldsymbol{\beta},\sigma^2|\mathbf{X},\mathbf{y})$ defined in Equation 23.22 as well as the prior distributions $p(\boldsymbol{\beta}|\,\sigma^2)$ and $p(\sigma^2)$, we have the necessary components to obtain the joint posterior distribution of the model parameters given the data. Specifically, the joint posterior distribution of the parameters $\boldsymbol{\beta}$ and σ^2 is given as

$$p\left(\boldsymbol{\beta},\sigma^2|y,\mathbf{X}\right)\propto L\left(\boldsymbol{\beta},\sigma^2|\mathbf{X},\mathbf{y}\right)\times p\left(\boldsymbol{\beta}|\sigma^2\right)\times p\left(\sigma^2\right), \quad (23.26)$$

which, after some algebra, yields

$$p\left(\boldsymbol{\beta},\sigma^2|\mathbf{y},\mathbf{X}\right)\propto\sigma^{-n-a}\exp\left[-\frac{1}{2\sigma^2}\begin{pmatrix}\hat{\sigma}^2(n-k)+(\boldsymbol{\beta}-\hat{\beta})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\hat{\beta})\\+2b+(\boldsymbol{\beta}-\mathbf{B})'(\boldsymbol{\beta}-\mathbf{B})\end{pmatrix}\right] \quad (23.27)$$

which has the form of a multivariate normal distribution.

Data

This example presents a small sample size Bayesian regression analysis. Data from PISA (OECD 2003b, 2010) and *Education at a Glance* (EAG; OECD 2003a, 2009) were obtained from 25 countries both in 2000 and 2009. Variables used in the regression were teacher salary, class size, GDP per capita, and aggregated reading performance. Teacher salary and per capita GDP were retrieved from EAG, and class size and reading performance were obtained from PISA 2000 and 2009 results (OECD 2003a,b, 2010). Per capita GDP was measured in equivalent U.S. dollars converted using a purchasing power parity formula. Teacher salary was measured relative to national income, that is, a ratio of the teacher salary after 15 years of experience (minimum training) to per capita GDP.

For this example, we use the Mplus software program (Muthén and Muthén 2012). Our focus on Mplus is based on the fact that it has a very general framework that allows for the specification of Bayesian models.

Results

The top two panels of Figure 23.1 show the trace plots and autocorrelation plots for the regression coefficient relating country-aggregated teacher salary to country-level reading competency. Remaining plots are available upon request. An inspection of the trace plots and autocorrelation plots show evidence of convergence. Identical results were found for remaining model parameters. Moreover, the scale reduction factor is approximately 1.0 for both cases, indicating excellent convergence of the two chains.

Table 23.1 shows the results of the Bayesian regression analysis using noninformative and informative priors as described above. Specifically, for the noninformative priors case, a normal prior was chosen for the regression coefficients, with a mean of zero and variance of 10^{10}, and a noninformative inverse-gamma prior was chosen for the residual variance. For the informative case, a normal prior was again chosen for the regression coefficients with means based on the results of a conventional regression using the PISA 2000 data. The prior variances of the regression coefficients were obtained by squaring the standard errors obtained from the conventional regression. This example demonstrates the use of prior information based on a previous ILSA cycle.

An inspection of Table 23.1 reveals, as expected, that the posterior standard deviations and PPIs are wider for the noninformative case than the informative case, reflecting our uncertainty regarding the model parameters. The top panel of Figure 23.2 shows the posterior density plots for slope of teacher salary on reading, where we can clearly see the differences between the noninformative and informative cases.

The bottom panel of Figure 23.2 shows the posterior predictive checking scatterplot under the noninformative and informative analyses. Recall that this plot is used to aid in evaluating the model's goodness-of-fit, with the proportion of observations above the 45-degree line

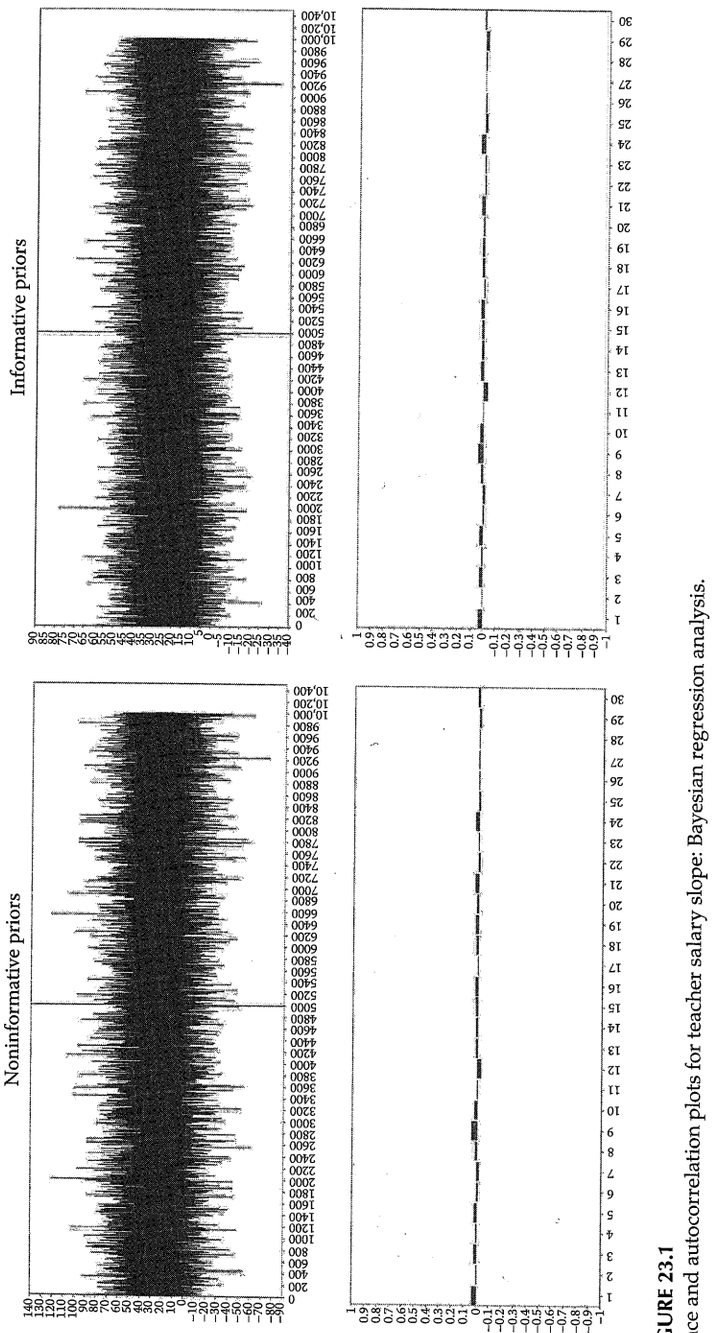

FIGURE 23.1
Trace and autocorrelation plots for teacher salary slope: Bayesian regression analysis.

TABLE 23.1

Between-Country Bayesian Regression Estimates Using PISA 2009 and EAG Data

Parameter	MAP	SD	*p*-value	95% PPI
Noninformative Priors				
READING on GDP Per cap	0.003	0.001	0.001	0.001, 0.004
READING on teacher salary	23.94	20.34	0.12	−17.44, 63.32
READING on class size	−0.84	1.44	0.50	−2.84, 2.88
Informative Priors (PISA 2000)				
READING on GDP per cap	0.003	0.001	0.000	0.002, 0.004
READING on teacher salary	22.32	12.27	0.04	−2.67, 46.10
READING on class size	−0.58	0.97	0.58	−1.68, 2.14

Note: MAP, maximum *a posteriori*; SD, posterior standard deviation; *p*-value is one-tailed.

corresponding to the Bayesian *p*-value. We find that the regression model with informative priors shows slightly better fit than the model with noninformative priors. Finally, the DIC values for the noninformative and informative priors cases is 243.64 and 241.448, respectively, favoring the model with informative priors.

We conclude this example by noting that in the case of small sample sizes (here, 25 countries) the influence of the priors is fairly noticeable. Our results regarding the precision of the estimates based on using information from the PISA 2000 cycle are considerably different than if we had not utilized this information.

EXAMPLE 23.2: BAYESIAN MULTILEVEL MODELING

A common feature of ILSA data collection is that students are nested in higher organizational units such as classrooms and/or schools. Indeed, in many instances, the substantive problem concerns specifically an understanding of the role that classrooms or school characteristics play in predicting an outcome of interest. For example, the PISA structure deliberately samples schools (within a country) and then takes an age-based sample of 15-year-olds within sampled schools. Such data collection plans are generically referred to as *clustered sampling designs*. Data from such clustered sampling designs are then collected at both levels for the purposes of understanding each level separately, but also to understand the inputs and processes of student- and school-level variables as they predict both school- and student-level outputs.

It is probably without exaggeration to say that one of the most important contributions to the empirical analysis of data arising from clustered sampling designs such as PISA has been the development of multilevel models. Important contributions to the theory of multilevel modeling can be found in Raudenbush and Bryk (2002) and references therein. In this example, we present Bayesian multilevel modeling.

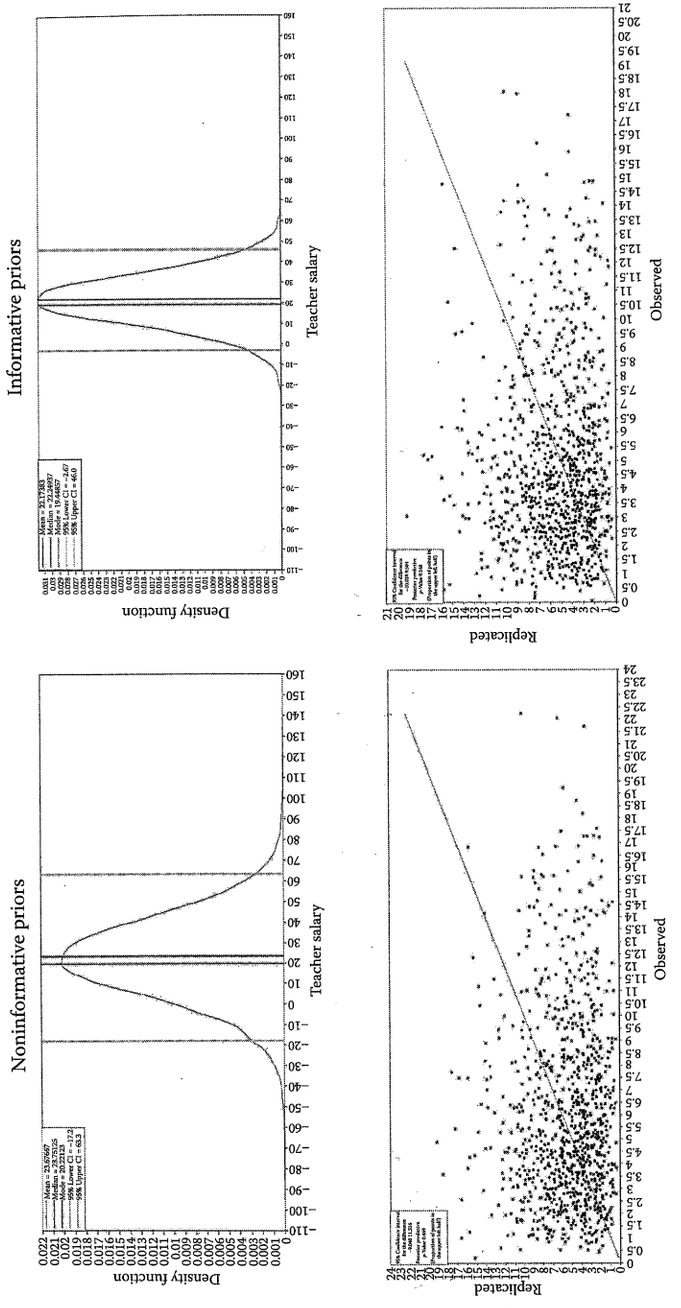

FIGURE 23.2
Posterior density and posterior probability scatter plots for teacher salary slope: Bayesian regression analysis.

Model

Perhaps the most basic multilevel model is the random effects analysis of variance model. As a simple example consider whether there are differences among G schools ($g = 1,2,...,G$) on the outcome of student achievement y obtained from n students ($i = 1,2,...,n$). In this example, it is assumed that the G schools are a random sample from a population of schools.* The model can be written as a two-level, random effects ANOVA model as follows. Let

$$y_{ig} = \beta_g + \epsilon_{ij}, \tag{23.28}$$

where y_{ig} is an achievement score for student i and school g, β_g is the school random effect, and ϵ_{ig} is an error term with homoskedastic variance σ^2. The model for the school random effect can be written as

$$\beta_g = \mu + \delta_g, \tag{23.29}$$

where μ is a grand mean and δ_g is an error term with homoskedastic variance ω^2 that picks up the school effect over and above the grand mean. Inserting Equation 23.29 into Equation 23.28 yields

$$y_{ig} = \mu + \delta_g + \varepsilon_{ig}, \tag{23.30}$$

which expresses the outcome y_{ig} in terms of an overall grand mean μ, a between-school effect δ_g, and a within-school effect ε_{ig}.

Recall that a fully Bayesian perspective requires specifying prior distributions on all model parameters. For the model in Equation 23.30, we first specify the distribution of the achievement outcome y_{ig} given the school effect δ_g and the within-school variance σ^2. Specifically

$$y_{ig} \left| \delta_g, \sigma^2 \sim N\left(\delta_g, \sigma^2\right) \right. \tag{23.31}$$

Given that parameters are assumed to be random in the Bayesian context, we next specify the prior distribution on the remaining model parameters. For this model, we specify normal conjugate priors for the school effects δ_g and the overall grand mean μ—viz.

$$\delta_g \left| \mu, \omega^2 \sim N(\mu, \omega^2), \right. \tag{23.32}$$

$$\mu \sim N\left(b_0, B_0\right), \tag{23.33}$$

* In many large-scale studies of schooling, the schools themselves may be obtained from a complex sampling scheme. However, we will stay with the simple example of random sampling.

where b_0 and B_0 are the mean and variance hyperparameters on μ that are assumed to be *fixed* and *known*. For the within-school and between-school variances, we specify the conjugate inverse-gamma priors—viz.

$$\sigma^2 \sim \text{inverse-gamma}(v_0/2, v_0\sigma_0^2/2), \tag{23.34}$$

$$\omega^2 \sim \text{inverse-gamma}(k_0/2, k_0\omega_0^2/2), \tag{23.35}$$

where v and k are degrees of freedom and σ_0^2 and ω_0^2 are hyperparameter values (Gelman et al. 2003).

To see how this specification fits into a Bayesian hierarchical model, note that we can arrange all of the parameters of the random-effects ANOVA model into a vector $\boldsymbol{\theta}$ and write the prior density as

$$p(\boldsymbol{\theta}) = p(\delta_1, \delta_2, \ldots, \delta_G, \mu, \sigma^2, \omega^2), \tag{23.36}$$

where under the assumption of exchangeability of the school effects δ_g we obtain (see, e.g., Jackman 2009)

$$p(\boldsymbol{\theta}) = \prod_{g=1}^{G} p\left(\delta_g \middle| \mu, \omega^2\right) p(\mu) p\left(\sigma^2\right) p\left(\omega^2\right) \tag{23.37}$$

Slopes and Intercepts as Outcomes Model

In the simple, random-effects ANOVA model, exchangeability warrants the existence of prior distributions on the school means β_g. We noted that a condition where exchangeability might not hold is if we are in possession of some knowledge about the schools, for example, if some are public schools and others are private schools. In this case, exchangeability across the entire set of schools is not likely to hold, and instead we must invoke *conditional exchangeability*. That is, we might be willing to accept exchangeability within school types. Our knowledge of school type, therefore, warrants the specification of a more general multilevel model that specifies the school means as a function of school-level characteristics. The addition of covariates at the student and school levels was first discussed in the educational context by Burstein (1980) and later developed by Raudenbush and Bryk (2002).*

* In this section, we have made the distinction between random-effects ANOVA and multilevel models. This is simply a matter of nomenclature. One could consider all of these models as a special case of hierarchical Bayesian models.

Data

This example presents a Bayesian multilevel regression analysis based on an unweighted sample of 5000 15-year-old students in the United States who were administered PISA 2009 (OECD 2012). The first plausible value of reading performance served as a dependent variable and was regressed on a set of student-level and school-level predictors. Student-level predictors included student background variables—specifically, gender (gender), immigrant status (native), language that they use (slang; coded 1 if test language is the same as language at home, 0 otherwise), and a measure of the student's ESCS. In addition, measures of student engagement and strategies in reading were included as predictors; specifically, enjoyment of reading (joyread), diversity in reading (divread), memorization strategies (memor), elaboration strategies (elab), control strategies (cstrat), student relationship with teachers (studrel), disciplinary climate (disclima), and class size (clsiz). A random slope is specified for the regression of reading performance on ESCS.

School-level predictors included school background variables; that is, school average socioeconomic background (xescs); school size (schsize) and square of school size (schsize2); city (coded 1 for both a small city and large city; 0 otherwise); and rural (coded 1 for a village, hamlet, rural area, or a small town; 0 otherwise). In addition, measures of school climate and policies were included; that is, school average student relationship with teachers (xstudrel), school average disciplinary climate (xdisclim), student behavior (studbeha), teacher behavior (teacbeha), student selection policies (selsch), transferring policy (transfer), school autonomy (respires, respcur), private school (private), school policies on assessment (stdtest, assmon, asscomp), school average language-learning time (xlmins), school average science-learning time (xsmins), school average mathematics-learning time (xmmins), shortage in staff (tcshort), and educational material (scmatedu). The slope of reading performance on ESCS is regressed on school average student relations with teachers (xstudrel) school disciplinary climate (xdisclim).

Method

Two multilevel regressions were conducted in a manner similar to Example 23.1. First, a Bayesian multilevel regression was conducted with weakly informative priors. This analysis assumes a normal distribution for the regression coefficients with a mean of zero and variance of 10^{10}. Weakly informative inverse-gamma priors were chosen for the residual variances. Thus, although the prior has a mode, the precision is so small as to be effectively noninformative. Second, a Bayesian multilevel regression was conducted with informative priors obtained from a conventional multilevel regression analysis of the PISA 2000 data, a manner similar to the regression analysis in Example 1 except that weakly informative inverse-gamma priors were used for the residual variances.

Results

The analysis used the Gibbs sampler as implemented in Mplus with two chains, 100,000 iterations with 50,000 burn-in and a thinning interval of

50. The default in Mplus is to discard half of the total number of iterations as burn-in. Thus, summary statistics on the model parameters are based on 1,000 draws from the posterior distribution generated via the Gibbs sampler. Figure 23.3 shows the trace plots and autocorrelation plots for both the noninformative and informative priors cases focusing on the random slope of reading performance on joyread. An inspection of these plots shows evidence of convergence. Moreover, in each case, the PSRF is very close to 1.0, indicating that the two chains have converged. Plots for all remaining parameters also indicate convergence. These plots are available upon request.

Tables 23.2 and 23.3 present selected results for the multilevel model based on noninformative and informative priors, respectively. We concentrate on predictors that are not part of the sampling design, thus these estimates are conditioned not only on the predictors in Table 23.2, but also on design variables not shown. The influence of priors can be clearly seen when examining the random slope regression. Recall the SLOPE refers to the regression coefficient relating reading performance to parental social and cultural status. For the noninformative priors case, the MAP estimate of SLOPE regressed on XSTUDREL is 0.42 (s.d. = 6.86) with a one-tailed p-value of 0.36. The 95% PPI ranges from –11.22 to 16.04. By contrast, the results of the informative case show a MAP estimate of –1.41 (s.d. = 5.15) with a one-tailed p-value of 0.75 and a 95% PPI ranging from –6.77 to 13.53. PPC plots and DIC values are not presently available in Mplus for Bayesian multilevel models.

EXAMPLE 23.3: BAYESIAN CONFIRMATORY FACTOR ANALYSIS

Model

Recent discussions of Bayesian confirmatory factor analysis and its extension to Bayesian structural equation modeling can be found in Kaplan and Depaoli (2012b), Lee (2007), and Muthén and Asparouhov (2012a). Following the general notation originally provided by Jöreskog (1969), write the confirmatory factor analysis model as

$$\mathbf{y} = \boldsymbol{\alpha} + \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}, \tag{23.38}$$

where $\mathbf{y}$ is a vector of manifest variables, $\boldsymbol{\alpha}$ is a vector of measurement intercepts, $\boldsymbol{\Lambda}$ is a factor loading matrix, $\boldsymbol{\eta}$ is a vector of latent variables, and $\boldsymbol{\varepsilon}$ is a vector of uniquenesses with covariance matrix $\boldsymbol{\Psi}$, typically specified to be diagonal. Under conventional assumptions (see, e.g., Kaplan 2009), we obtain the model expressed in terms of the population covariance matrix $\boldsymbol{\Sigma}$ as

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}, \tag{23.39}$$

where $\boldsymbol{\Phi}$ is the covariance matrix of the common factors. The distinction between the confirmatory factor analysis (CFA) model in Equation 23.38

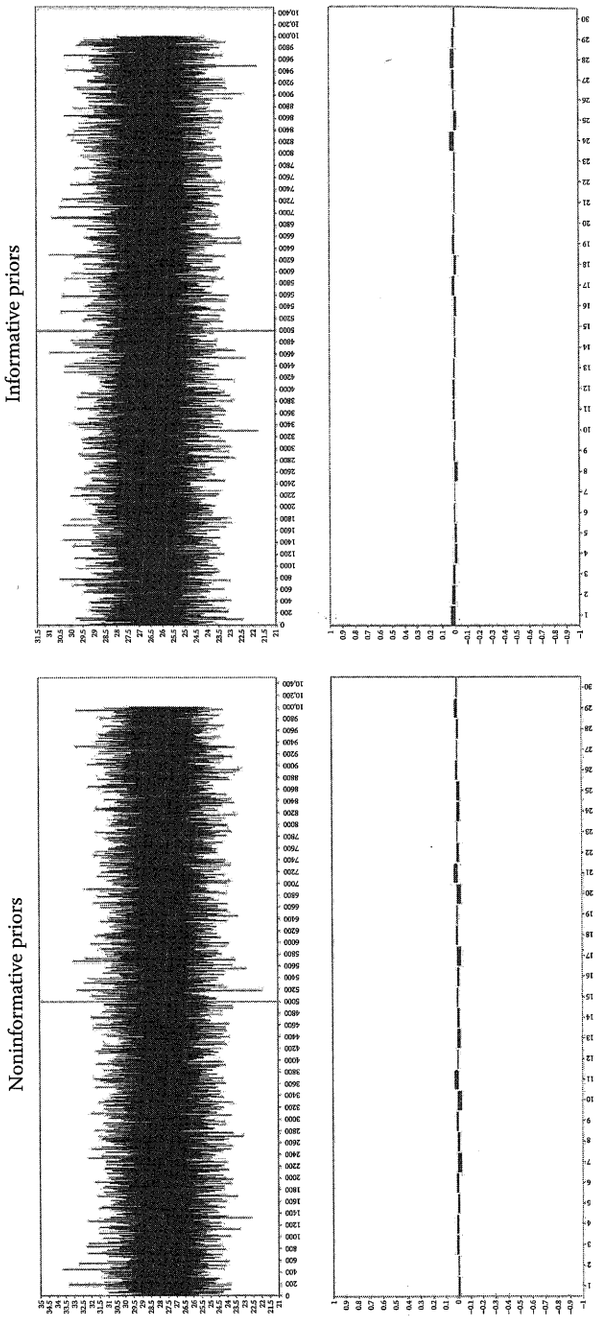

FIGURE 23.3
Trace and autocorrelation plots for joy-of-reading slope: Bayesian hierarchical linear model.

TABLE 23.2

Multilevel Bayesian Regression Estimates with Noninformative Priors

Parameter	MAP	SD	*p*-value	95% PPI
Within Level				
READING ON JOYREAD	27.28	1.36	0.000	25.10, 30.45
READING ON DIVREAD	−3.43	1.34	0.004	−6.16, −0.93
READING ON MEMOR	−13.08	1.46	0.000	−18.28, −12.55
READING ON ELAB	−12.38	1.39	0.000	−14.60, −9.17
READING ON CSTRAT	21.70	1.61	0.000	18.56, 24.83
READING ON STUDREL	1.31	1.24	0.11	−0.94, 3.89
READING ON DISCLIMA	6.40	1.31	0.000	4.87, 10.01
READING ON CLSIZ	0.71	0.19	0.001	0.24, 0.97
Between Level				
SLOPE ON XSTUDREL	0.42	6.90	0.36	−11.22, 16.04
SLOPE ON XDISCLIM	5.69	6.30	0.09	−3.95, 20.67
READING ON XSTUDREL	−13.40	10.50	0.11	−33.71, 7.60
READING ON XDISCLIM	24.24	9.41	0.01	12.34, 49.59
READING ON STUDBEHA	0.83	3.91	0.05	−1.21, 14.19
READING ON TEACBEHA	0.28	3.77	0.54	−7.57, 7.20
READING ON SELSCH	3.88	5.28	0.23	−6.45, 14.28
READING ON TRANSFER	1.03	6.89	0.42	−12.03, 14.91
READING ON RESPRES	−6.02	3.20	0.34	−7.73, 5.13
READING ON RESPCUR	5.55	2.24	0.30	−3.15, 5.54
READING ON PRIVATE	36.74	24.57	0.06	−10.00, 86.16
READING ON STDTEST	−73.08	33.01	0.04	−124.72, 5.26
READING ON ASSMON	−4.15	18.53	0.43	−39.61, 33.58
READING ON ASSCOMP	12.67	8.26	0.18	−8.74, 23.27
READING ON XLMINS	−0.03	0.09	0.76	−0.11, 0.24
READING ON XSMINS	−0.14	0.08	0.09	−0.19, 0.04
READING ON XMMINS	0.15	0.09	0.49	−0.18, 0.18
READING ON TCSHORT	−2.39	2.98	0.23	−7.90, 3.65
READING ON SCMATEDU	2.38	2.32	0.78	−6.36, 2.76

Note: MAP, maximum *a posteriori*; SD, posterior standard deviation; *p*-value is one-tailed.

and exploratory factor analysis typically lies in the number and location of restrictions placed in the factor loading matrix $\boldsymbol{\Lambda}$ (see, e.g., Kaplan 2009).

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. For this model, let $\boldsymbol{\theta}_{\text{norm}} = \{\boldsymbol{\alpha}, \boldsymbol{\Lambda}\}$ be the set of free model parameters that are

TABLE 23.3

Multilevel Bayesian Regression Estimates with Informative Priors Based on PISA 2000

Parameter	MAP	SD	*p*-value	95% PPI
Within Level				
READING ON JOYREAD	28.44	1.13	0.000	24.21, 28.65
READING ON DIVREAD	−2.48	1.13	0.007	−3.87, 0.53
READING ON MEMOR	−16.47	1.26	0.000	−18.94, −13.99
READING ON ELAB	−10.25	1.19	0.000	−14.07, −9.44
READING ON CSTRAT	20.36	1.35	0.000	19.74, 25.01
READING ON STUDREL	1.02	1.06	0.04	−0.18, 3.97
READING ON DISCLIMA	5.44	1.10	0.000	2.29, 6.57
READING ON CLSIZ	0.65	0.18	0.000	0.36, 1.07
Between Level				
SLOPE ON XSTUDREL	−1.41	5.15	0.75	−6.77, 13.53
SLOPE ON XDISCLIM	9.12	4.80	0.23	−5.85, 12.78
READING ON XSTUDREL	−2.28	7.00	0.54	−12.82, 14.29
READING ON XDISCLIM	16.37	6.75	0.01	2.34, 28.71
READING ON STUDBEHA	2.14	3.03	0.14	−2.69, 9.13
READING ON TEACBEHA	5.34	2.88	0.38	−4.74, 6.47
READING ON SELSCH	−4.03	3.70	0.21	−10.32, 4.21
READING ON TRANSFER	−5.37	5.56	0.43	−12.05, 9.81
READING ON RESPRES	−2.43	2.13	0.39	−4.84, 3.57
READING ON RESPCUR	−1.10	1.72	0.33	−4.07, 2.70
READING ON PRIVATE	20.18	13.85	0.02	1.62, 56.47
READING ON STDTEST	−16.46	10.55	0.03	−40.12, 1.12
READING ON ASSMON	−0.58	9.68	0.72	−13.56, 24.43
READING ON ASSCOMP	6.90	6.66	0.35	−10.51, 15.50
READING ON XLMINS	0.07	0.06	0.20	−0.06, 0.17
READING ON XSMINS	−0.02	0.05	0.39	−0.11, 0.08
READING ON XMMINS	−0.05	0.06	0.15	−0.18, 0.06
READING ON TCSHORT	−0.16	1.99	0.32	−4.85, 2.98
READING ON SCMATEDU	−4.67	1.95	0.15	−5.93, 1.86

Note: MAP, maximum *a posteriori*; SD, posterior standard deviation; *p*-value is one-tailed.

assumed to follow a normal distribution and let $\boldsymbol{\theta}_{IW} = \{\boldsymbol{\Phi}, \boldsymbol{\Psi}\}$ be the set of free model parameters that are assumed to follow an inverse-Wishart distribution. Thus

$$\boldsymbol{\theta}_{\text{norm}} \sim N(\boldsymbol{\mu}, \boldsymbol{\Omega}), \tag{23.40}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$ are the mean and variance hyperparameters, respectively, of the normal prior. The uniqueness covariance matrix $\boldsymbol{\Psi}$ is assumed to follow an inverse-Wishart distribution. Specifically

$$\boldsymbol{\theta}_{IW} \sim IW(\mathbf{R}, \delta), \tag{23.41}$$

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 inverse-Wishart distribution.

Data

This example is based on a reanalysis of a confirmatory factor analysis described in the OECD technical report (OECD 2012). In the report, the confirmatory factor analysis was employed to construct two indices indicating teacher and student behavioral problems (TEACBEHA and STUDBEHA), using a weighted sample of students from the OECD countries. For this example, we used an unweighted sample of 165 school principals in the United States who participated in PISA 2009. The principals were administered a questionnaire asking to what extent student learning is hindered by student or teacher behavioral problems. Each item has the following four categories: not at all, very little, to some extent, and a lot.

The CFA model in this example was specified to have two factors, which are teacher and student behavioral problems. The factor related to teacher behavioral problems contains the following seven items: teachers' low expectation of students (SC17Q01), poor student–teacher relations (SC17Q03), teachers not meeting individual students' needs (SC17Q05), teacher absenteeism (SC17Q06), staff resisting change (SC17Q09), teachers being too strict with students (SC17Q11), and students not being encouraged to achieve their full potential (SC17Q13). The second factor relating to student behavioral problems contains the following six items: student absenteeism (SC17Q02), disruption of classes by students (SC17Q04), students skipping classes (SC17Q07), students lacking respect for teachers (SC17Q08), student use of alcohol or illegal drugs (SC17Q10), and students intimidating or bullying other students (SC17Q12).

Results

The analysis used the Gibbs sampler as implemented in Mplus with two chains, 100,000 iterations with 50,000 burn-in and a thinning interval of 50. Thus, summary statistics on the model parameters are based on 1000 draws from the posterior distribution generated via the Gibbs sampler. Figure 23.4 presents the trace plots and autocorrelation plots for both the noninformative and informative cases. The plots show evidence of convergence, and the PSRF (not shown) is very close to 1.0.

Selected results of the CFA model for the noninformative (upper panel) and informative (lower panel) cases is displayed in Table 23.4. For the noninformative case, a normal prior was chosen for the factor loadings, with a mean of zero and variance of 10^{10}, and a noninformative inverse-gamma prior was chosen for the factor variances and unique variances. For the informative priors case, priors on the factor loadings were based

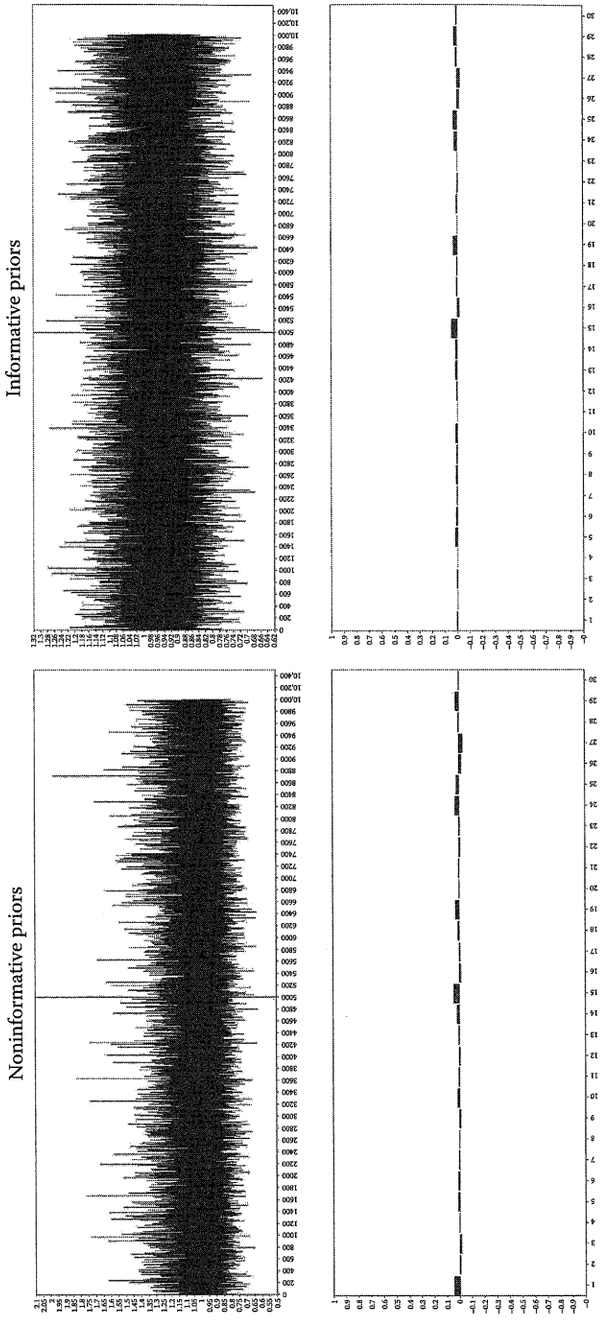

FIGURE 23.4
Trace and autocorrelation plots for factor loading three: Bayesian confirmatory factor analysis.

TABLE 23.4

Selected Bayesian CFA Estimates with Noninformative and Informative Priors

Parameter	MAP	SD	*p*-value	95% PPI
Noninformative Priors				
Loadings: TEABEHA by				
SC17Q03	0.99	0.13	0.00	0.78, 1.31
SC17Q05	0.94	0.13	0.00	0.70, 1.20
SC17Q06	0.68	0.12	0.00	0.48, 0.95
SC17Q09	0.87	0.14	0.00	0.73, 1.28
SC17Q11	0.56	0.11	0.00	0.31, 0.73
SC17Q13	0.96	0.14	0.00	0.74, 1.27
Loadings: STUDBEHA by				
SC17Q04	0.85	0.13	0.00	0.69, 1.18
SC17Q07	0.98	0.15	0.00	0.82, 1.41
SC17Q08	0.93	0.14	0.00	0.78, 1.30
SC17Q10	0.59	0.11	0.00	0.34, 0.78
SC17Q12	0.56	0.09	0.00	0.42, 0.78
TEABEHA with				
STUDBEHA	0.18	0.04	0.00	0.12, 0.27
Informative Priors				
Loadings: TEABEHA by				
SC17Q03	1.00	0.08	0.00	0.80, 1.11
SC17Q05	1.07	0.09	0.00	0.82, 1.17
SC17Q06	0.77	0.09	0.00	0.58, 0.94
SC17Q09	0.97	0.10	0.00	0.82, 1.19
SC17Q11	0.64	0.08	0.00	0.40, 0.72
SC17Q13	1.06	0.10	0.00	0.84, 1.21
Loadings: STUDBEHA by				
SC17Q04	0.92	0.09	0.00	0.75, 1.11
SC17Q07	1.12	0.12	0.00	0.95, 1.42
SC17Q08	1.02	0.10	0.00	0.86, 1.24
SC17Q10	0.73	0.10	0.00	0.48, 0.88
SC17Q12	0.67	0.08	0.00	0.51, 0.83
TEABEHA with				
STUDBEHA	0.15	0.03	0.00	0.12, 0.23

Note: MAP, maximum *a posteriori*; SD, posterior standard deviation.

on a previous factor analysis of the PISA 2000 data in a manner similar to Example 23.1. Weakly informative inverse-gamma priors were chosen for the factor variances and unique variances.

As expected, including informative priors based on a conventional CFA of the PISA 2000 data yields smaller posterior standard deviations and narrower 95% PPIs when compared to the noninformative priors case. An inspection of the posterior density plot for one of the loadings (TEABEHA by SC17Q06) in the upper panel of Figure 23.5 shows the

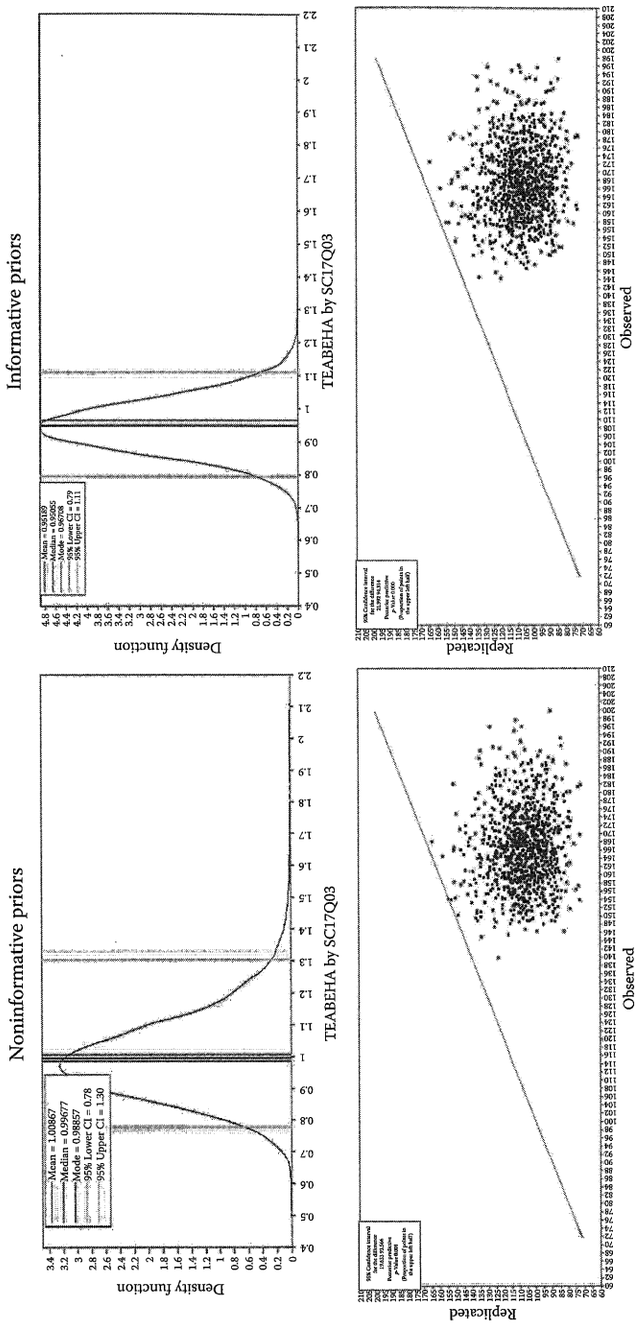

FIGURE 23.5
Posterior density and posterior probability scatter plots for loading 3: Bayesian confirmatory factor analysis.

difference between the noninformative case and informative case with respect to the shape of the posterior density.

An inspection of the lower panel of Figure 23.5 shows the PPC scatterplot for noninformative priors and informative priors cases. We see that virtually all of the likelihood ratio chi-square values fall below the 45-degree line, indicating poor model fit to the posterior replicated data. As in conventional confirmatory factor analysis, lack of model fit may be due to the restrictions placed on the factor loading matrix in line with the theory that there are two factors underlying these data. In the interest of space, we do not modify this model; however, see Muthén and Asparouhov (2012a) for an example of model modification in the Bayesian CFA context. Finally, the DIC values for the noninformative and informative priors cases for the CFA model are 3505.32 and 3503.50, slightly favoring the CFA model with informative priors.

Conclusion

The purpose of this chapter was to discuss and illustrate the Bayesian approach to the analysis of ILSA data. The chapter provided a brief overview of the elements of Bayesian inference along with an example illustrating the implementation of a multilevel model from a Bayesian perspective.

It is worth asking why one would choose to adopt the Bayesian framework for the analysis of ILSA data—particularly when, in large samples, it can often provide results that are very close to that of frequentist approaches such as maximum likelihood. 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. It must be noted that despite the similarities in the results, the interpretations are completely different. First, from the Bayesian perspective, parameters are viewed as random and unknown, reflecting our uncertainty about unknown quantities, with probability serving as the language of uncertainty. This is in contrast to the frequentist approach, which views parameters as fixed and unknown. Second, the Bayesian perspective evaluates the quality of a substantive model in terms of posterior prediction, with competing models judged in terms of their support within the data. This is in contrast to conventional null hypothesis testing with its focus on assessing a hypothesis that is known *a priori* not to be true. Finally, the summary of the posterior distribution of the model parameters reflects our current or "updated" knowledge about the parameters of interest, and this updated knowledge should be incorporated in future studies in the form of new priors. No such notion of "updating" knowledge exists in the frequentist framework, and each analysis is treated as though nothing was learned from previous studies.

Clearly, then, the critical difference relates to reflecting uncertainty via the specification of the prior distribution.

A fair question to ask of the Bayesian approach centers on how priors should be obtained. 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" (O'Hagan et al. 2006; Abbas et al. 2008, 2010). In some cases, elicitation of prior knowledge can be obtained from experts and/or key stakeholders (however, see Muthén and Asparouhov [2012b] for a discussion of the dangers of using informative priors favored by a researcher). In the context of ILSAs, however, we have demonstrated how informative prior information can be gleaned directly from previous waves of the same ILSA—in our case, PISA 2000—and incorporated into a Bayesian model specification. Alternative elicitations from different cycles of the same ILSA and even different ILSAs can be directly compared via Bayesian model selection measures, such as use of the DIC or Bayes factors.

To summarize, we believe that conventional frequentist statistical modeling cannot exploit all that can be learned from ILSAs such as PISA. In contrast, we believe that Bayesian inference, with its focus on formally combining current data with previous research, can provide a methodological framework for the *evolutionary* development of knowledge about the inputs, processes, and outcomes of schooling. The practical benefits of the Bayesian approach for international educational research 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 R305D110001 to the University of Wisconsin—Madison. The opinions expressed are those of the authors and do not necessarily represent views of the institute or the U.S. Department of Education.

References

Abbas, A. E., Budescu, D. V., and Gu, Y. 2010. Assessing joint distributions with isoprobability contours. *Management Science*, 56, 997–1011.

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

Berger, J. 2006. The case for objective Bayesian analysis. *Bayesian Analysis*, 3, 385–402.
Brooks, S. P. and Gelman, A. 1998. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7, 434–455.
Burstein, L. 1980. The analysis of multilevel data in educational research and evaluation. *Review of Research in Education*, 8, 158–233.
de Finetti, B. 1974. *Theory of Probability*, vols. 1 and 2. New York: John Wiley and Sons.
Fox, J. 2008. *Applied Regression Analysis and Generalized Linear Models*, 2nd edition. Newbury Park , CA: Sage.
Gelman, A. 1996. Inference and monitoring convergence. In: W. R. Gilks, S. Richardson, and D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice*. New York: Chapman & Hall, pp. 131–143.
Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. 2003. *Bayesian Data Analysis*. 2nd edition. London: Chapman and Hall.
Gelman, A. and Rubin, D. B. 1992a. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457–511.
Gelman, A. and 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, and A. F. M. Smith (Eds.), *Bayesian Statistics 4*. Oxford: Oxford University Press, pp. 625–631.
Geman, S. and 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, and A. F. M. Smith (Eds.), *Bayesian Statistics 4*. Oxford: Oxford University Press.
Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (Eds.). 1996. *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.
Gill, J. 2002. *Bayesian Methods: A Social and Behavioral Sciences Approach*. London: Chapman and Hall/CRC.
Heidelberger, P. and Welch, P. 1983. Simulation run length control in the presence of an initial transient. *Operations Research*, 31, 1109–1144.
Howson, C. and Urbach, P. 2006. *Scientific Reasoning: The Bayesian Approach*. 3rd edition. Chicago: Open Court.
Jackman, S. 2009. *Bayesian Analysis for the Social Sciences*. New York: John Wiley.
Jöreskog, K. G. 1969. A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34, 183–202.
Kaplan, D. 2009. *Structural Equation Modeling: Foundations and Extensions*. 2nd edition. Newbury Park , CA: Sage Publications.
Kaplan, D. and Depaoli, S. 2012a. Bayesian statistical methods. In T. D. Little (Ed.), *Oxford Handbook of Quantitative Methods*. Oxford: Oxford University Press.
Kaplan, D. and Depaoli, S. 2012b. Bayesian structural equation modeling. In: R. Hoyle (Ed.), *Handbook of Structural Equation Modeling*. Guilford Publishing, Inc, pp. 650–673.
Kass, R. E. and Raftery, A. E. 1995. Bayes factors. *Journal of the American Statistical Association*, 90, 773–795.
Kim, J.-S. and Bolt, D. M. 2007. Estimating item response theory models using Markov chain Monte Carlo methods. *Educational Measurement: Issues and Practice*, 26, 38–51.
Lee, S.-Y. 2007. *Structural Equation Modeling: A Bayesian Approach*. New York: Wiley.
Lynch, S. M. 2007. *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. New York: Springer.

Martin, M. O., Mullis, I. V. S., and Foy, P. 2008. *TIMSS 2007 International Science Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades*. Chestnut Hill, MA: Boston College.
Mullis, I. V. S., Martin, M. O., Kennedy, A. M., and Foy, P. 2008. *PIRLS 2006 International Science Report: IEA's Progress In International Reading Literacy Study in Primary Schools in 40 Countries*. Chestnut Hill, MA: Boston College.
Muthén, B. and Asparouhov, T. 2012a. Bayesian SEM: A more flexible representation of substantive theory. *Psychological Methods*, 17, 313–335.
Muthén, B. and Asparouhov, T. 2012b. Rejoinder: Mastering a new method. *Psychological Methods*, 17(3), 346–353. doi: 10.1037/a0029214.
Muthén, L. K. and Muthén, B. 2012. *Mplus: Statistical Analysis with Latent Variables*. Los Angeles: Muthén & Muthén.
OECD. 2003a. *Education at a Glance 2003: OECD Indicators*. Paris: OECD.
OECD. 2003b. *Literacy Skills for the World of Tomorrow: Further Results from PISA 2000*. Paris: Author.
OECD. 2009. *Education at a Glance 2009: OECD Indicators*. Paris: OECD.
OECD. 2010. *Results: What Makes a School Successful? Resources, Policies and Practices (Volume IV)*. Paris: Author.
OECD. 2012. *PISA 2009 Technical Report*. Paris: Author.
O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., and Rakow, T. 2006. *Uncertain Judgements: Eliciting Experts' Probabilities*. West Sussex, England: Wiley.
Press, S. J. 2003. *Subjective and Objective Bayesian Statistics: Principles, Models, and Applications*. 2nd edition. New York: Wiley.
Raftery, A. E. 1995. Bayesian model selection in social research (with discussion). In: P. V. Marsden (Ed.), *Sociological Methodology*. Vol. 25. New York: Blackwell, pp. 111–196.
Raftery, A. E. and Lewis, S. M. 1992. How many iterations in the Gibbs sampler? In: J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Eds.), *Bayesian Statistics 4*. Oxford: Oxford University Press, pp. 763–773.
Raudenbush, S. W. and Bryk, A. S. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods*. 2nd edition. Thousands Oaks, CA: Sage Publications.
Royall, R. 1997. *Statistical Evidence: A Likelihood Paradigm*. New York: Chapman and Hall.
Sinharay, S. 2004. Experiences with Markov chain Monte Carlo convergence assessment in two psychometric examples. *Journal of Educational and Behavioral Statistics*, 29, 461–488.
Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Linde, A. van der. 2002. Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 64, 583–639.
Wiseman, A. W. 2010. *The Impact of International Achievement Studies on National Education Policy Making*. Bingley, UK: Emerald, Publishing.