Journal of Educational and Behavioral Statistics
Summer 1996, Vol. 21, No. 2, pp. 131–167

Bayesian Analysis in Applications of Hierarchical Models: Issues and Methods

Michael H. Seltzer
University of California, Los Angeles
Wing Hung Wong
Chinese University of Hong Kong
Anthony S. Bryk
University of Chicago

Key words: *hierarchical models, fully Bayesian analysis, Gibbs sampling, multivariate* t *priors*

In applications of hierarchical models (HMs), a potential weakness of empirical Bayes estimation approaches is that they do not to take into account uncertainty in the estimation of the variance components (see, e.g., Dempster, 1987). One possible solution entails employing a fully Bayesian approach, which involves specifying a prior probability distribution for the variance components and then integrating over the variance components as well as other unknowns in the HM to obtain a marginal posterior distribution of interest (see, e.g., Draper, 1995; Rubin, 1981). Though the required integrations are often exceedingly complex, Markov-chain Monte Carlo techniques (e.g., the Gibbs sampler) provide a viable means of obtaining marginal posteriors of interest in many complex settings. In this article, we fully generalize the Gibbs sampling algorithms presented in Seltzer (1993) to a broad range of settings in which vectors of random regression parameters in the HM (e.g., school means and slopes) are assumed multivariate normally or multivariate t *distributed across groups. Through analyses of the data from an innovative mathematics curriculum, we examine when and why it becomes important to employ a fully Bayesian approach and discuss the need to study the sensitivity of results to alternative prior distributional assumptions for the variance components and for the random regression parameters.*

This work was made possible in part by a Spencer Dissertation Year Fellowship awarded to Michael Seltzer. We wish to thank the University of Chicago School Mathematics Project for allowing us to use the data from the Transition Mathematics Field Study to illustrate the issues and methods discussed in this article. The development and evaluation of the Transition Mathematics curriculum was supported by grants from the Amoco Foundation and the Carnegie Foundation. We would also like to thank the associate editor, an anonymous reviewer, Don Rubin, and Stephen Raudenbush for very helpful comments on earlier drafts of this article.

A form of the hierarchical model (HM) that has been used extensively in analyzing continuous outcomes is a two-level formulation in which vectors of Level 1 regression parameters (e.g., school means and slopes) are assumed multivariate normally (MVN) distributed across Level 2 units (e.g., schools). The two-level, MVN formulation has been used in growth curve analyses (e.g., Laird & Ware, 1982; Strenio, Weisberg, & Bryk, 1983), reliability and predictive validity studies (e.g., Novick, Jackson, & Thayer, 1971; Braun, Jones, Rubin, & Thayer, 1983), studies of school effects (e.g., Aitkin & Longford, 1986; de Leeuw & Kreft, 1986; V. Lee & Bryk, 1989; Raudenbush & Bryk, 1986), and analyses of the effects of educational interventions (e.g., Raffe, 1991).

In its general form, Level 1 of the two-level HM consists of a regression model in which the observations (y_{ij}, $i = 1, \ldots, n_j$) nested within each of J Level 2 units ($j = 1, \ldots, J$) are modeled as a function of R predictor variables ($X_{1ij}, X_{2ij}, \ldots, X_{Rij}$):

$$y_{ij} = \beta_{j0} + \beta_{j1}X_{1ij} + \beta_{j2}X_{2ij} + \cdots + \beta_{jR}X_{Rij} + e_{ij}, \tag{1}$$

where the β_{jr} ($r = 0, \ldots, R$) are unknown regression parameters, and the e_{ij} are residuals assumed normally distributed with mean 0 and variance σ^2.

All $R + 1$ regression parameters in the above model could be treated as varying across Level 2 groups (i.e., random), or the variation in 1 or more parameters might be constrained to be 0 (see, for example, Aitkin and Longford's [1986] discussion of "constant slopes" models). Thus it is useful to reexpress the Level 1 model in the following manner:

$$\mathbf{y}_j = \mathbf{X}_j\boldsymbol{\beta}_j + \mathbf{X}_{j+}\boldsymbol{\gamma}_+ + \mathbf{e}_j, \tag{2}$$

where $\mathbf{y}_j$ is an $n_j \times 1$ vector of observations for group j, $\boldsymbol{\beta}_j$ is a $P \times 1$ vector consisting of those regression parameters treated as varying across groups (termed *random regression parameters* in this article), and $\boldsymbol{\gamma}_+$ is an $F \times 1$ vector comprised of those parameters that are treated as nonvarying or fixed. Values on the predictor variables connected with each of these sets of coefficients are contained in the matrices $\mathbf{X}_j$ and $\mathbf{X}_{j+}$, which are dimensioned $n_j \times P$ and $n_j \times F$, respectively. Finally, $\mathbf{e}_j$ is an $n_j \times 1$ vector of residuals assumed normally distributed with mean vector $\mathbf{0}$ and variance $\sigma^2\mathbf{I}_{n_j}$.

At Level 2, the random regression parameters are modeled as a function of Level 2 characteristics ($\mathbf{W}_j$):

$$\boldsymbol{\beta}_j \mid \boldsymbol{\gamma}, \mathbf{T} \sim N_P(\mathbf{W}_j\boldsymbol{\gamma}, \mathbf{T}), \tag{3}$$

where $\boldsymbol{\gamma}$ is a $K \times 1$ vector of fixed effects relating differences in the magnitude of the random regression parameters to differences in Level 2 characteristics,

and $\mathbf{T}$ is a $P \times P$ matrix capturing the variance and covariance of the elements of $\boldsymbol{\beta}_j$ about $\mathbf{W}_j\boldsymbol{\gamma}$, a vector of expected values conditional on $\mathbf{W}_j$.

Thus, in multisite evaluation studies, for example, interest might center on program effects for particular sites (e.g., elements of the $\boldsymbol{\beta}_j$) and on elements of $\boldsymbol{\gamma}$ relating differences in site characteristics to differences in the magnitude of within-site program effects. However, as detailed in the following section, problems can arise in utilizing standard, empirical Bayes (EB) approaches to estimation and inference for $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}_j$, particularly when the number of Level 2 units in a sample is small. In this article, the use of a fully Bayesian approach in addressing these problems is discussed and implemented via a Markov-chain Monte Carlo (MCMC) technique termed the *Gibbs sampler* (Gelfand & Smith, 1990).

Seltzer (1993) discusses the use of the Gibbs sampler in simple HM settings in which the Level 1 model contains a single regression parameter, for example, $y_{ij} = \beta_j + \epsilon_{ij}$, or $y_{ij} = \beta_j X_{ij} + \epsilon_{ij}$. Algorithms are presented for settings in which the β_j are assumed normally distributed or, so that sensitivity analyses can be conducted (see below), univariate t distributed. In this article, we generalize this work and present Gibbs sampling algorithms that can be used to calculate marginal posteriors of interest in HMs in which vectors of Level 1 regression parameters are assumed MVN or multivariate t (MVT) distributed. The algorithms that we outline can be applied in a broad range of settings commonly encountered in practice—that is, settings in which each of the regression parameters specified in the HM is treated as varying across groups, or in which the variation in one or more regression parameters is constrained to be 0.

EB Estimation of $\boldsymbol{\gamma}$ *and* $\boldsymbol{\beta}_j$

To convey problems connected with EB estimation for the HM, and to introduce key distributional forms that appear in the Gibbs sampling algorithms detailed below, we give the conditional posterior distributions for $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}_j$ given $\mathbf{T}$, σ^2, and the data ($\mathbf{y}$), based on an HM in which each of the regression parameters in the Level 1 model is treated as random (Laird & Ware, 1982; Lindley & Smith, 1972). Treating Equation 3 as the prior distribution for $\boldsymbol{\beta}_j$, $j = 1, \ldots, J$, where $\mathbf{W}_j\boldsymbol{\gamma}$ and $\mathbf{T}$ are, respectively, the prior mean and variance-covariance matrix of $\boldsymbol{\beta}_j$, and assuming a uniform prior for $\boldsymbol{\gamma}$, we have

$$\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{T}, \sigma^2 \sim N_K(\boldsymbol{\gamma}^*, \mathbf{D}^*), \tag{4}$$

where

$$\boldsymbol{\gamma}^* = \left[\sum_{j=1}^{J} \mathbf{W}_j'(\mathbf{V}_j + \mathbf{T})^{-1}\mathbf{W}_j\right]^{-1} \sum_{j=1}^{J} \mathbf{W}_j'(\mathbf{V}_j + \mathbf{T})^{-1}\hat{\boldsymbol{\beta}}_j,$$

and

$$\mathbf{D}^* = \left[\sum_{j=1}^{J} \mathbf{W}_j'(\mathbf{V}_j + \mathbf{T})^{-1}\mathbf{W}_j \right]^{-1},$$

where $\hat{\boldsymbol{\beta}}_j = (\mathbf{X}'_j\mathbf{X}_j)^{-1}\,\mathbf{X}'_j\mathbf{y}_j$ is the ordinary least squares (OLS) estimate of $\boldsymbol{\beta}_j$ with sampling variance $\mathbf{V}_j = \sigma^2(\mathbf{X}'_j\mathbf{X}_j)^{-1}$, and $\boldsymbol{\gamma}^*$ is the familiar generalized least squares (GLS) estimator for $\boldsymbol{\gamma}$ discussed in numerous references (see, for example, Raudenbush, 1988).

For the $\boldsymbol{\beta}_j$ we have

$$\boldsymbol{\beta}_j \mid \mathbf{y}, \mathbf{T}, \sigma^2 \sim N_P(\boldsymbol{\beta}_j^*, \mathbf{V}_j^*), \tag{5}$$

where

$$\boldsymbol{\beta}_j^* = \boldsymbol{\Lambda}_j\hat{\boldsymbol{\beta}}_j + (\mathbf{I}_P - \boldsymbol{\Lambda}_j)\mathbf{W}_j\boldsymbol{\gamma}^*,$$

with

$$\boldsymbol{\Lambda}_j = \mathbf{T}(\mathbf{V}_j + \mathbf{T})^{-1} = (\mathbf{V}_j^{-1} + \mathbf{T}^{-1})^{-1}\mathbf{V}_j^{-1},$$

and

$$\mathbf{I}_P - \boldsymbol{\Lambda}_j = (\mathbf{V}_j^{-1} + \mathbf{T}^{-1})^{-1}\mathbf{T}^{-1},$$

and with $\boldsymbol{\gamma}^*$, $\boldsymbol{\beta}_j$, and $\mathbf{V}_j$ defined as above. It can be seen that $\boldsymbol{\beta}_j^*$ is the well-known composite (or shrinkage) estimator based on normally distributed data and a normal prior (see, for example, Lindley & Smith, 1972; Laird & Ware, 1982; and Strenio et al., 1983). As the precision of $\hat{\boldsymbol{\beta}}_j$ decreases, the amount of weight placed on the estimated prior mean $\mathbf{W}_j\boldsymbol{\gamma}^*$ increases. The conditional posterior variance of $\boldsymbol{\beta}_j$ (see Raudenbush, 1988) has the following form:

$$\mathbf{V}_j^* = \boldsymbol{\Lambda}_j\mathbf{V}_j + (\mathbf{I}_P - \boldsymbol{\Lambda}_j)\mathbf{S}_j(\mathbf{I}_P - \boldsymbol{\Lambda}_j)', \tag{6}$$

with

$$\mathbf{S}_j = \mathbf{W}_j\mathbf{D}^*\mathbf{W}_j'.$$

In employing an EB approach to estimation and inference for $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}_j$, maximum likelihood (ML) estimates of the variance components are obtained using one of a number of iterative techniques (e.g., EM or Fisher scoring),

and point estimates and intervals for $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}_j$ would be based on the conditional posterior distributions defined in Equations 4 and 5 setting $\mathbf{T}$ and σ^2 equal to their ML estimates. Thus, we have

$$\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{T} = \mathbf{T}_{\text{ml}}, \sigma^2 = \sigma^2_{\text{ml}} \sim N_K(\boldsymbol{\gamma}^*_{\text{ml}}, \mathbf{D}^*_{\text{ml}}) \tag{7}$$

and

$$\boldsymbol{\beta}_j \mid \mathbf{y}, \mathbf{T} = \mathbf{T}_{\text{ml}}, \sigma^2 = \sigma^2_{\text{ml}} \sim N_P(\boldsymbol{\beta}^*_{j\text{ml}}, \mathbf{V}^*_{j\text{ml}}), \tag{8}$$

where $\boldsymbol{\gamma}^*_{\text{ml}}$ and $\boldsymbol{\beta}^*_{j\text{ml}}$ are EB estimates of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}_j$, respectively. (For discussions of ML estimation of the variance components in the HM, see Dempster, Rubin, & Tsutakawa, 1981; de Leeuw & Kreft, 1986; Goldstein, 1986; and Longford, 1987.)

Problems Connected With the EB Approach and Potential Solutions via a Fully Bayesian Approach

A clear concern that arises in using the EB approach is that the posterior variances ($\mathbf{D}^*_{\text{ml}}$; $\mathbf{V}^*_{j\text{ml}}$) and intervals that one obtains (Equations 7 and 8) do not take into account the uncertainty connected with using estimates of the variance components in place of their true values (Dempster, 1987). This is especially problematic in settings where the number of Level 2 units (J) in a sample is small, since the amount of information available for estimating $\mathbf{T}$ depends in large part on J. Hence, the intervals that we obtain in such situations may be misleadingly small.

To help clarify this issue, it is useful to consider the one-sample problem where we have $y_i \sim N(\mu, \sigma^2)$, $i = 1, \ldots, n$. Treating the sample variance (s^2) as the true value for σ^2 and, in turn, basing intervals for μ on the conditional posterior $p(\mu \mid \mathbf{y}, \sigma^2 = s^2)$ yields a 95% interval for μ of $\bar{Y} \pm 1.96(s/\sqrt{n})$. We know that as the number of observations for estimating σ^2 becomes small, such an approach becomes problematic. But while t procedures are available to us in such settings to take into account uncertainty concerning the estimation of σ^2 from the data, standard adjustments of this kind are unavailable in all but the simplest HMs (see, for example, Kirk's [1982] treatment of balanced hierarchical designs).

Morris (1983), Kackar and Harville (1984), and Laird and Louis (1987) provide methods for adjusting intervals for random regression parameters in cases where the Level 1 model consists of a single, unknown regression parameter. However, Laird and Louis (1989) remark that in models where vectors of regression parameters are assumed to vary across Level 2 units—as in, for example, Equation 3—adjusting the posterior distributions defined in Equations 7 and 8 to reflect uncertainty concerning $\mathbf{T}$ and σ^2 becomes extremely difficult.

An additional problem stemming from small numbers of Level 2 units has been discussed by Rubin (1981): namely, point estimates of $\boldsymbol{\beta}_j$ based on ML estimates of the variance components may constitute poor summaries of the data. This arises from the fact that when J is small, the likelihood functions for Level 2 variance parameters will tend to exhibit a high degree of asymmetry; hence, the resulting ML estimate of $\mathbf{T}$ used to weight the data and the prior mean (Equation 5) may be unrepresentative of plausible values for $\mathbf{T}$ (see also Draper, 1995).

A potential solution to addressing problems connected with small numbers of Level 2 units is to employ a fully Bayesian approach. In this approach, priors are placed on the variance components, and then we attempt to integrate over the variance components as well as other unknowns in the model to obtain the marginal posterior distribution of a fixed effect or random regression parameter of interest. In the case of the one-sample problem, for example, the fully Bayesian approach involves placing priors on μ and σ^2 and then integrating out σ^2 from the joint posterior $p(\mu, \sigma^2 \mid \mathbf{y})$ to obtain the marginal posterior $p(\mu|\mathbf{y})$ (see, for example, Box & Tiao, 1973, chap. 2).

Rubin (1981) carried out a fully Bayesian analysis in his effort to draw inferences concerning the magnitude of the effect of coaching on student performance on the SAT in eight parallel experiments ($j = 1, \ldots, 8$). In that application, the Level 1 model consisted of a single, unknown regression parameter (i.e., a coaching effect) assumed normally distributed at Level 2, and marginal posteriors could be calculated via numerical integration techniques or, as illustrated by Rubin, multiple-imputation-based methods. However, in cases where vectors of regression parameters ($\boldsymbol{\beta}_j$) are assumed MVN distributed, the use of a fully Bayesian approach has been impeded due to the extreme difficulty of the required integrations. One of the aims of this article, as noted above, is to present a general Gibbs sampling formulation for settings of this kind.

In utilizing a fully Bayesian approach, one must be aware that the answers one obtains concerning parameters of interest (e.g., posterior intervals) may be sensitive to choice of priors for the variance components, especially when sample sizes (J) are small. In the next section, we discuss several priors that appear in the literature (e.g., Jeffreys' prior), paying particular attention to their strengths and weaknesses. Whereas Seltzer (1993), for example, places uniform priors on the variance components (see also Rubin, 1981), we illustrate the important analytic practice of recalculating marginal posteriors for parameters of interest under alternative specifications of priors for the variance components.

In addition to studying the sensitivity of results to choice of priors for the variance components, it is also prudent to study the sensitivity of results to assumptions of normality at Level 2. In particular, this involves reanalyses of the data under heavy-tailed distributional assumptions. With respect to drawing inferences concerning $\boldsymbol{\gamma}$, Seltzer (1993) has noted that as a least

squares estimator, $\boldsymbol{\gamma}^*$ (Equation 4) is sensitive to extreme $\hat{\boldsymbol{\beta}}$ (see also West, 1984). For settings in which the Level 1 model consists of a single, unknown regression parameter (β_j), Seltzer (1991, 1993) and Carlin (1992) show how the Gibbs sampler can be used to calculate marginal posterior distributions of fixed effects of interest in HMs in which the β_j are assumed univariate t distributed, thereby yielding results that are robust to outlying Level 2 units (e.g., an unusually effective school). In this article we extend these algorithms to a range of settings in which vectors of regression parameters are assumed MVT distributed at Level 2.

While Seltzer (1993) and Carlin (1992) focus on the sensitivity of results for fixed effects to assumptions of normality, an additional problem connected with the Level 2 normality assumption centers on the EB estimator $\boldsymbol{\beta}^*_{j\,\mathrm{ml}}$. Specifically, shrinkage can be severe for elements of $\hat{\boldsymbol{\beta}}_j$ that are far from the estimated prior mean $\mathbf{W}_j\boldsymbol{\gamma}^*_{\mathrm{ml}}$. Thus, Efron and Morris (1971, 1972) warn that while EB estimators may perform well in terms of estimation of an entire ensemble of random regression parameters (e.g., a set of J scalars β_j, $j = 1, \ldots, J$) as measured by squared error loss ($\Sigma_{j=1}^{J}\ (\beta^*_{j\,\mathrm{ml}} - \beta_j)^2$), they may perform quite poorly in connection with estimating those β_j that differ substantially from the estimated prior mean; that is, the degree of bias in such cases is potentially high. Thus, while substantive interest often focuses on, for example, unusually effective schools, we run the risk of seriously misestimating such effects when employing the compromise estimator that is obtained under the assumption of normal priors. The limited translation rules developed by Efron and Morris (1971, 1972), which constrain the amount by which the EB and OLS estimates may differ, are an attempt to remedy this problem (for an alternative view on limited translation rules, see Rubin, 1980). However, as Dempster (1983) notes, limited translation estimators "are effectively ad hoc devices corresponding roughly to the use of Bayesian priors for [Level 1 parameters] with longer than normal tails" (p. 57). Thus, from a Bayesian perspective, it would be prudent to conduct analyses under the assumption of heavy-tailed (e.g., t) priors, as well as under the assumption of normal priors.

In the next section, we examine possible choices of priors for the variance components and then outline Gibbs sampling algorithms for calculating marginal posteriors of interest in the HM under MVN and MVT distributional assumptions for the $\boldsymbol{\beta}_j$. We then utilize the Gibbs sampler as well as a standard EB approach in analyses of the data from an evaluation of an innovative mathematics curriculum called Transition Mathematics. A key aim of these analyses is to illustrate and explain differences in location and posterior variance that can arise for random regression parameters and fixed effects of interest as we move from an EB approach to a fully Bayesian approach. Attention is drawn to situations in which it becomes particularly important to employ a fully Bayesian approach.

Implementing the Fully Bayesian Approach in Applications of the HM Under MVN and MVT Level 2 Assumptions

Choosing Priors for the Variance Components

To conduct fully Bayesian analyses, prior probability distributions must be chosen for the variance components in the HM (i.e., σ^2 and $\mathbf{T}$). There are several alternatives, and their advantages and disadvantages require some discussion.

Two of the possible choices of a prior for σ^2 discussed in the literature are a uniform prior—that is, $p(\sigma^2) \propto k$, where k is a constant—and Jeffreys' prior—that is, $p(\sigma^2) \propto 1/\sigma^2$. In contrast to the uniform prior, Jeffreys' prior assigns small weight to increasingly large values of σ^2 and places large amounts of weight on values of σ^2 approaching 0; in fact, it is readily seen that $1/\sigma^2$ becomes infinitely large for values of $\sigma^2 \approx 0$.

Whereas both of these priors are improper, an inverse chi-square prior of the following form provides us with a proper prior:

$$p(\sigma^2) \propto (\sigma^2)^{-[(\nu_1/2)+1]} \exp\left[-\frac{S_1}{2\sigma^2}\right], \tag{9}$$

with degrees of freedom $\nu_1 > 0$ and scale parameter $S_1 > 0$. Unlike the uniform prior and Jeffreys' prior, Equation 9 defines a class of distributions that are unimodal (see Novick & Jackson, 1974). For $1 < \nu_1 < 4$, the inverse chi-square distribution has infinite variance and hence would provide us with a prior that is weak relative to the information provided by the data (see, for example, Gelfand, Hills, Racine-Poon, & Smith, 1990). In addition, for small ν_1, the inverse chi-square prior exhibits strong, positive skew, and so, in contrast to the uniform prior, prior probabilities gradually decrease as values of σ^2 become arbitrarily large. Furthermore, unlike Jeffreys' prior, the density function depicted in Equation 9 does not become infinitely large as σ^2 approaches 0. Thus, an inverse chi-square prior with small degrees of freedom constitutes an attractive alternative to the uniform prior and Jeffreys' prior (see P. Lee, 1989).

In addressing the issue of possible priors for $\mathbf{T}$, it is helpful to begin with a discussion of settings in which we have a single variance component in the Level 2 model, that is, τ. As in the case of priors for σ^2, three potential choices are $p(\tau) \propto k$, $p(\tau) \propto 1/\tau$, and an inverse chi-square prior with degrees of freedom $\nu_2 > 0$ and scale parameter $S_2 > 0$. (Whereas a subscript of 1 was used to denote the degrees of freedom and scale parameter for the prior distribution of the Level 1 variance parameter, the subscript 2 will be used in the case of the prior for the Level 2 variance component(s).)

Morris (1983), Lindley (1983), and DuMouchel and Waternaux (1992) point out that the use of Jeffreys' prior for τ is problematic. Even if only a tiny fraction of the mass of the likelihood $l(\tau|\mathbf{y})$ lies near $\tau = 0$, the fact that

$1/\tau$ becomes infinitely large for values of $\tau \approx 0$ will result in a posterior distribution for τ that is a spike at $\tau = 0$. The problem is that $\tau = 0$ always "has a non-zero likelihood value due to the presence of sampling variation when $\sigma > 0$" (DuMouchel & Waternaux, 1992, p. 339). Hence, DuMouchel and Waternaux discourage the use of Jeffreys' prior. They also note that because of possible numerical inaccuracies of MCMC methods such as the Gibbs sampler, analysts might sometimes miss this problem (see, for example, Geyer, 1992).

An appealing feature of a uniform prior for τ is that the modes of the posterior and likelihood will be identical, since the likelihood is reproduced as the posterior, that is, $p(\tau|\mathbf{y}) \propto l(\tau|\mathbf{y})$. A potential drawback, however, is that prior probabilities attached to τ do not decrease as values of τ become arbitrarily large (DuMouchel & Waternaux, 1992); as a result, the posterior intervals that we obtain for fixed effects and random regression parameters of interest may tend to be somewhat conservative.

A viable alternative to the uniform prior would be a weak inverse chi-square prior (i.e., $1 < \nu_2 < 4$) whose mode is approximately equal to the mode of $l(\tau|\mathbf{y})$. A weak prior of this kind would slightly favor values of τ near the mode of $l(\tau|\mathbf{y})$, thereby giving some support to the information provided by the data (see Gelfand et al., 1990). An advantage over the uniform prior is that this prior would downweight arbitrarily large values of τ (see Note 1). Though not explored in this article, an additional option would be to specify a more informative inverse chi-square prior with mode and spread reflecting, for example, information based on previous empirical work in a particular area (DuMouchel & Waternaux, 1992).

In settings where a matrix of variances and covariances is specified in the Level 2 model ($\mathbf{T}$), the inverse chi-square distribution generalizes to an inverse Wishart distribution:

$$p(\mathbf{T}) \propto |\mathbf{T}|^{-[(\nu_2+P+1)/2]} \exp\left[-\frac{1}{2} \text{tr } \mathbf{T}^{-1}\mathbf{S}_2\right], \tag{10}$$

with degrees of freedom $\nu_2 \geq P$ and where $\mathbf{S}_2$ is a positive definite symmetric scale matrix (see Zellner, 1971). Similar to the inverse chi-square prior, setting ν_2 to a small value provides us with a prior that is weak relative to the information provided by the data, and, as will be seen, $\mathbf{S}_2$ can be chosen so that the prior lends some support to the information provided by the data regarding $\mathbf{T}$.

As we will see later in an illustrative example, rather than simply choosing one prior for the variance components, it is prudent to study the sensitivity of one's results to reasonable, alternative specifications of the prior, particularly when sample sizes are small.

Using the Gibbs Sampler to Obtain Marginal Posteriors of Interest Under MVN Level 2 Distributional Assumptions

Implementation of the Gibbs sampler in simple HM settings in which the Level 1 model consists of a single unknown regression parameter which is assumed univariate normally distributed is presented in various sources including Morris (1987), Gelfand & Smith (1990), and Seltzer (1991, 1993). Kasim (1994) outlines a Gibbs sampling algorithm for an HM consisting of two unknown regression parameters in the Level 1 model: an intercept assumed univariate normal ($P = 1$) and a slope whose variance is constrained to be 0 ($F = 1$) (see Equation 2 above). Also focusing on Level 1 models consisting of two unknown regression parameters, Gelfand et al. (1990) and Seltzer (1991) illustrate the use of the Gibbs sampler in settings in which the regression parameters are assumed MVN distributed ($P = 2$) (see also Raudenbush, Cheong, & Fotiu, 1995). In this section, we integrate and generalize these lines of work. We outline an algorithm, based on the general form of the Level 1 model depicted in Equations 1 and 2, that can be applied in settings in which all regression parameters are assumed MVN distributed, or in which the variation in one or more regression parameters is constrained to be 0.

To implement the fully Bayesian approach, we must place priors on all unknowns. For the $P \times 1$ vector of random regression parameters in the Level 1 model (see Equation 2) we have: $\boldsymbol{\beta}_j|\boldsymbol{\gamma}, \mathbf{T} \sim N_P(\mathbf{W}_j\boldsymbol{\gamma}, \mathbf{T})$. For the $F \times 1$ vector of regression parameters treated as nonvarying across groups ($\boldsymbol{\gamma}_+$ in Equation 2), we assume a uniform prior. A uniform prior is also assumed for the $K \times 1$ vector of fixed effects $\boldsymbol{\gamma}$ relating differences in Level 2 characteristics ($\mathbf{W}_j$) to differences in $\boldsymbol{\beta}_j$. For the variance components, we assume an inverse chi-square prior for σ^2 and an inverse Wishart prior for $\mathbf{T}$. Thus, combining the data and these priors, the joint posterior distribution of $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \ldots, \boldsymbol{\beta}'_J)'$, $\boldsymbol{\gamma}$, $\boldsymbol{\gamma}_+$, $\mathbf{T}$, and σ^2 is as follows:

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\gamma}_+, \mathbf{T}, \sigma^2|\mathbf{y}) \propto \prod_{j=1}^{J} p(\mathbf{y}_j|\boldsymbol{\beta}_j, \boldsymbol{\gamma}_+, \sigma^2) \prod_{j=1}^{J} p(\boldsymbol{\beta}_j|\boldsymbol{\gamma}, \mathbf{T}) \times p(\boldsymbol{\gamma}_+)\, p(\boldsymbol{\gamma})\, p(\sigma^2)\, p(\mathbf{T}), \quad (11)$$

where

$$p(\mathbf{y}_j|\boldsymbol{\beta}_j, \boldsymbol{\gamma}_+, \sigma^2) \propto \prod_{i=1}^{n_j} (1/\sigma^2)^{1/2} \exp\left[-\frac{1}{2\sigma^2} (y_{ij} - [\mathbf{X}'_{ij}\boldsymbol{\beta}_j + \mathbf{X}'_{ij+}\boldsymbol{\gamma}_+)^2 \right] \quad (12)$$

and

$$p(\boldsymbol{\beta}_j|\boldsymbol{\gamma}, \mathbf{T}) \propto |\mathbf{T}^{-1}|^{1/2} \exp\left[-\frac{1}{2}(\boldsymbol{\beta}_j - \mathbf{W}_j\boldsymbol{\gamma})'\mathbf{T}^{-1}(\boldsymbol{\beta}_j - \mathbf{W}_j\boldsymbol{\gamma})\right], \quad (13)$$

and where $p(\sigma^2)$ and $p(\mathbf{T})$ are defined as in Equations 9 and 10, respectively.

Integrating over all unknowns in Equation 11 to obtain marginal posteriors of interest is forbiddingly complex—that is, the required integrations are intractable. As a viable alternative, however, we can use the Gibbs sampler to obtain draws from the joint posterior and then simulate the marginal posterior distribution of any parameter of interest via the empirical distribution of the values generated for that parameter. (For a thorough discussion of the Gibbs sampler, see Gelfand & Smith, 1990; see also Tanner & Wong, 1987; Morris, 1987; Zeger & Karim, 1991; Tanner, 1993; Casella & George, 1992; Gelman & Rubin, 1992; and Seltzer, 1993.)

The essential feature of the Gibbs sampler is that we subdivide the unknowns in the joint posterior in a way that makes it easy (or possible) to sample from the conditional posterior distribution of each group of unknowns given the other groups of unknowns and the data. As such, the algorithm we present entails sampling from the conditional posterior distributions for the following five groups of unknowns: σ^2, $\boldsymbol{\beta}$, $\mathbf{T}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\gamma}_+$. To obtain a sense of how the steps comprising the Gibbs sampler are derived, see the univariate normal formulation presented by Seltzer (1993) (it should be noted that the latter formulation gives results only for the case where uniform priors are placed on the variance components in the HM).

Step 1. $p(\sigma^2|\mathbf{y}, \boldsymbol{\beta}, \mathbf{T}, \boldsymbol{\gamma}, \boldsymbol{\gamma}_+)$:

$$1/\sigma^2 \sim \text{Gamma}\left[a = [(N + \nu_1)/2], b = \left[2 \Big/ \left(S_1 + \sum_{j=1}^{J}\sum_{i=1}^{n_j} e_{ij}^2\right)\right]\right], \quad (14)$$

where $e_{ij} = (y_{ij} - [\mathbf{X}'_{ij}\boldsymbol{\beta}_j + \mathbf{X}'_{ij+}\boldsymbol{\gamma}_+])$, $N = \sum_{j=1}^{J} n_j$, and ν_1 and S_1 are the degrees of freedom and scale parameters of the prior for σ^2 specified in Equation 9. Note that $E(1/\sigma^2) = ab$ and $\text{Var}(1/\sigma^2) = ab^2$. Under the prior $p(\sigma^2) \propto 1/\sigma^2$, we have $\nu_1 = 0$ and $S_1 = 0$; for $p(\sigma^2) \propto k$, we have $\nu_1 = -2$ and $S_1 = 0$.

When each of the regression parameters specified in the Level 1 model is treated as varying across groups, the residuals in Equation 14 (i.e., the e_{ij}) simply reduce to $e_{ij} = (y_{ij} - \mathbf{X}'_{ij}\boldsymbol{\beta}_j)$.

Step 2. $p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{T}, \boldsymbol{\gamma}, \boldsymbol{\gamma}_+, \sigma^2)$: With $\boldsymbol{\gamma}_+$ given, Equation 2 becomes

$$\mathbf{d}_j = \mathbf{X}_j\boldsymbol{\beta}_j + \mathbf{e}_j, \quad (15)$$

where $\mathbf{d}_j = \mathbf{y}_j - \mathbf{X}_{j+}\boldsymbol{\gamma}_+$. It can be seen that $\mathbf{d}_j$ is the vector of observations for group j adjusted for differences in the values of the predictors contained in $\mathbf{X}_{j+}$. Based on the data for group j and given $\boldsymbol{\gamma}_+$, the OLS estimate of $\boldsymbol{\beta}_j$ has the following form:

$$\hat{\boldsymbol{\beta}}_{(d)j} = (\mathbf{X}_j'\mathbf{X}_j)^{-1}\mathbf{X}_j'\mathbf{d}_j. \tag{16}$$

Combining the data and the prior for $\boldsymbol{\beta}_j$ gives rise to the following conditional posterior distribution for this step of the algorithm:

$$\boldsymbol{\beta}_j \sim N_P(\tilde{\boldsymbol{\beta}}_j, \tilde{\mathbf{V}}_j) \qquad (j = 1, \ldots, J), \tag{17}$$

where

$$\tilde{\boldsymbol{\beta}}_j = \boldsymbol{\Lambda}_j\hat{\boldsymbol{\beta}}_{(d)j} + (\mathbf{I}_P - \boldsymbol{\Lambda}_j)\mathbf{W}_j\boldsymbol{\gamma},$$

and

$$\tilde{\mathbf{V}}_j = \boldsymbol{\Lambda}_j\mathbf{V}_j = (\sigma^{-2}\mathbf{X}_j'\mathbf{X}_j + \mathbf{T}^{-1})^{-1},$$

with $\boldsymbol{\Lambda}_j = (\mathbf{V}_j^{-1} + \mathbf{T}^{-1})^{-1}\mathbf{V}_j^{-1}$ and $\mathbf{V}_j = \sigma^2(\mathbf{X}_j'\mathbf{X}_j)^{-1}$. Drawing from the work of Braun et al. (1983), when $\mathbf{X}_j'\mathbf{X}_j$ is singular for a particular group, we can reexpress $\tilde{\boldsymbol{\beta}}_j$ as follows:

$$\tilde{\boldsymbol{\beta}}_j = (\sigma^{-2}\mathbf{X}_j'\mathbf{X}_j + \mathbf{T}^{-1})^{-1}(\sigma^{-2}\mathbf{X}_j'\mathbf{d}_j + \mathbf{T}^{-1}\mathbf{W}_j\boldsymbol{\gamma}). \tag{18}$$

When all of the regression parameters in Equation 1 are treated as varying across groups, $\hat{\boldsymbol{\beta}}_{(d)j}$ in Equation 17 is simply replaced by $\hat{\boldsymbol{\beta}}_j = (\mathbf{X}_j'\mathbf{X}_j)^{-1}\mathbf{X}_j'\mathbf{y}_j$, and $\mathbf{y}_j$ replaces $\mathbf{d}_j$ in Equation 18.

Step 3. $p(\mathbf{T}|\mathbf{y}, \boldsymbol{\gamma}, \boldsymbol{\gamma}_+, \sigma^2, \boldsymbol{\beta})$:

$$\mathbf{T}^{-1} \sim \text{Wishart}_P\left[\mathbf{B} = \left[\mathbf{S}_2 + \sum_{j=1}^{J}\mathbf{U}_j\mathbf{U}_j'\right]^{-1}, \nu_w = (J + \nu_2)\right], \tag{19}$$

where $\mathbf{U}_j = (\boldsymbol{\beta}_j - \mathbf{W}_j\boldsymbol{\gamma})$, $\mathbf{B}$ is a $P \times P$ scale matrix (see, for example, Box & Tiao, 1973, p. 427), ν_w represents the degrees of freedom of the Wishart distribution, and ν_2 and $\mathbf{S}_2$ are, respectively, the degrees of freedom and scale matrix of the prior for $\mathbf{T}$ specified in Equation 10. Algorithms presented by Odell and Feiveson (1966) and Smith and Hocking (1972) provide an easy means of sampling from Wishart distributions. Assuming a uniform prior for $\mathbf{T}$, we have $\mathbf{S}_2 = \mathbf{0}$ and $\nu_2 = -(P + 1)$.

In settings where only one of the regression parameters specified at Level

1 is assumed normally distributed, the conditional posterior distribution for the variance component connected with that parameter (i.e., τ) under the assumption of an inverse chi-square prior (see above) is as follows (cf. Seltzer, 1993):

$$1/\tau \sim \text{Gamma}\left[a = [(J + \nu_2)/2],\ b = \left[2 \Big/ \left(S_2 + \sum_{j=1}^{J} U_j^2\right)\right]\right], \quad (20)$$

where $U_j = \beta_j - \mathbf{W}_j\boldsymbol{\gamma}$.

Step 4. $p(\boldsymbol{\gamma}|\mathbf{y}, \boldsymbol{\gamma}_+, \sigma^2, \boldsymbol{\beta}, \mathbf{T})$:

$$\boldsymbol{\gamma} \sim N_K(\tilde{\boldsymbol{\gamma}}, \tilde{\mathbf{D}}), \quad (21)$$

where

$$\tilde{\boldsymbol{\gamma}} = \left[\sum_{j=1}^{J} \mathbf{W}_j'\mathbf{T}^{-1}\mathbf{W}_j\right]^{-1} \sum_{j=1}^{J} \mathbf{W}_j'\mathbf{T}^{-1}\boldsymbol{\beta}_j,$$

and

$$\tilde{\mathbf{D}} = \left[\sum_{j=1}^{J} \mathbf{W}_j'\mathbf{T}^{-1}\mathbf{W}_j\right]^{-1}.$$

Step 5. $p(\boldsymbol{\gamma}_+|\mathbf{y}, \sigma^2, \boldsymbol{\beta}, \mathbf{T}, \boldsymbol{\gamma})$: With $\boldsymbol{\beta}_j$ known, Equation 2 becomes

$$\mathbf{d}_{j+} = \mathbf{X}_{j+}\boldsymbol{\gamma}_+ + \mathbf{e}_j, \quad (22)$$

where $\mathbf{d}_{j+} = \mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta}_j$. Thus, $\mathbf{d}_{j+}$ is the vector of observations for group j adjusted for differences in the values of the predictors contained in $\mathbf{X}_j$. Pooling the data for all groups yields the following conditional posterior for $\boldsymbol{\gamma}_+$:

$$\boldsymbol{\gamma}_+ \sim N_F(\tilde{\boldsymbol{\gamma}}_+, \tilde{\mathbf{D}}_+), \quad (23)$$

where

$$\tilde{\boldsymbol{\gamma}}_+ = \left[\sum_{j=1}^{J} \mathbf{X}_{j+}'\mathbf{X}_{j+}\right]^{-1} \sum_{j=1}^{J} \mathbf{X}_{j+}'\mathbf{d}_{j+},$$

and

$$\tilde{\mathbf{D}}_+ = \sigma^2\left[\sum_{j=1}^{J} \mathbf{X}'_{j+}\mathbf{X}_{j+}\right]^{-1}.$$

When all regression parameters specified at Level 1 are treated as random, Step 5 of the algorithm is, of course, omitted. (For a discussion of how nonrandomly varying slopes [Bryk & Raudenbush, 1992, pp. 21–22] are handled in this formulation, see the last paragraph of Appendix A.)

In implementing the algorithm, we choose starting values for the parameters in the model and then proceed through the steps of the algorithm, sampling once from each conditional posterior given the most recently sampled values for the other groups of unknowns. We continue cycling through the steps of the algorithm until convergence and then use values generated in subsequent cycles to simulate the marginal posterior distribution of any parameter of interest in the model. For details concerning implementation of the Gibbs sampler and monitoring convergence, see Gelman and Rubin (1992) and Seltzer (1993, pp. 213, 216, and 232). Before proceeding to the next section, note that an alternative formulation to that presented above could be developed within the generalized linear modeling framework outlined by Zeger and Karim (1991).

MVT Level 2 Distributional Assumptions

Seltzer (1991, 1993) and Carlin (1992) present Gibbs sampling algorithms for settings in which the Level 1 model consists of a single, unknown regression parameter (i.e., β_j) assumed t distributed. These formulations utilize an approach, outlined in Tanner and Wong (1987), based on the normal/gamma mixture representation of the t distribution. An MVT extension is illustrated in Seltzer (1991) (see also Racine-Poon, 1992). In this section, we build on this work and extend the algorithm presented above to settings in which vectors of regression parameters specified in the Level 1 model are assumed MVT distributed.[2] To help grasp the similarities in logic between the MVT and univariate t formulations, see Seltzer (1993).

We first write $\boldsymbol{\beta}_j - \mathbf{W}_j\boldsymbol{\gamma} = \mathbf{U}_j$ where $\mathbf{U}_j$ is assumed MVT (t_P) distributed with mean $\mathbf{0}$, scale $\mathbf{T}$, and ν degrees of freedom. (The nonsubscript symbol ν will be used in this article to refer to the degrees of freedom assumed for the distribution of the random regression parameters ($\boldsymbol{\beta}_j$) or random effects ($\mathbf{U}_j$).) Since an MVT distributed vector is equivalent to an MVN distributed vector scaled by a gamma distributed variate, we can write $\mathbf{U}_j$ in the following form:

$$(\mathbf{U}_j = \mathbf{R}'\mathbf{Z}_j q_j^{-1/2}) \sim t_P(\mathbf{0}, \mathbf{T}, \nu), \tag{24}$$

where $\mathbf{R}'\mathbf{R} = \mathbf{T}$, $\mathbf{Z}_j \sim N_P(\mathbf{0}, \mathbf{I}_P)$ and $q_j \sim \chi^2_\nu / \nu$.

The crux of the MVT algorithm presented below is that conditional on q_j, the Level 2 model becomes MVN with variance-covariance $\mathbf{T}q_j^{-1}$

$$\boldsymbol{\beta}_j | \mathbf{T}, \boldsymbol{\gamma}, q_j \sim N_P(\mathbf{W}_j\boldsymbol{\gamma}, \mathbf{T}q_j^{-1}), \tag{25}$$

thus resulting in a set of conditional posteriors that are easy to sample from. Before proceeding, we wish to point out that in contrast to Equation 3, where essentially we have $q_j = 1$ for each of J Level 2 units, the prior variances of the elements of $\boldsymbol{\beta}_j$ are large (i.e., prior information is weak) for those Level 2 units with small values of q_j.

Utilizing the MVN/gamma formulation of the multivariate t, the joint posterior is now as follows:

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\gamma}_+, \mathbf{T}, \sigma^2, \mathbf{q} | \mathbf{y}) \propto \prod_{j=1}^{J} p(\mathbf{y}_j | \boldsymbol{\beta}_j, \boldsymbol{\gamma}_+, \sigma^2) \prod_{j=1}^{J} p(\boldsymbol{\beta}_j | \boldsymbol{\gamma}, \mathbf{T}, q_j) \times \prod_{j=1}^{J} p(q_j)\, p(\boldsymbol{\gamma}_+)\, p(\boldsymbol{\gamma})\, p(\sigma^2)\, p(\mathbf{T}), \tag{26}$$

where

$$p(\boldsymbol{\beta}_j | \boldsymbol{\gamma}, \mathbf{T}, q_j) \propto |\mathbf{T}^{-1}|^{1/2} \exp\left[-\frac{1}{2} (\boldsymbol{\beta}_j - \mathbf{W}_j\boldsymbol{\gamma})'\mathbf{T}^{-1}q_j(\boldsymbol{\beta}_j - \mathbf{W}_j\boldsymbol{\gamma}) \right], \tag{27}$$

and

$$p(q_j) \propto q_j^{(\nu/2-1)} \exp\left[-\frac{\nu q_j}{2} \right], \tag{28}$$

where $\mathbf{q} = (q_1, \ldots, q_J)'$ constitutes a sixth group of unknowns, and $p(\mathbf{y}_j|\boldsymbol{\beta}_j, \boldsymbol{\gamma}_+, \sigma^2)$ and the priors for $\boldsymbol{\gamma}_+$, $\boldsymbol{\gamma}$, σ^2, and $\mathbf{T}$ are defined as in the MVN formulation presented in the previous section. The steps of a Gibbs sampling algorithm for this model are as follows.

Step 1. $p(\sigma^2|\mathbf{y}, \boldsymbol{\beta}, \mathbf{T}, \boldsymbol{\gamma}, \boldsymbol{\gamma}_+, \mathbf{q})$: This step of the algorithm is identical to Equation 14.

Step 2. $p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{T}, \boldsymbol{\gamma}, \boldsymbol{\gamma}_+, \mathbf{q}, \sigma^2)$: This step is similar in form to Equation 17, but with $\mathbf{T}q_j^{-1}$ replacing $\mathbf{T}$. Hence, as q_j decreases (i.e., as the prior variance of $\boldsymbol{\beta}$ increases), the amount of weight placed on the data in Equation 17 increases.

Step 3. $p(\mathbf{T}|\mathbf{y}, \boldsymbol{\gamma}, \boldsymbol{\gamma}_+, \sigma^2, \boldsymbol{\beta})$:

$$\mathbf{T}^{-1} \sim \text{Wishart}_P\left[\mathbf{B} = \left[\mathbf{S}_2 + \sum_{j=1}^{J} q_j \mathbf{U}_j \mathbf{U}_j'\right]^{-1}, \nu_w = (J + \nu_2)\right]. \quad (29)$$

Note that those Level 2 units with small values of q_j receive small amounts of weight in the calculation of the scale matrix $\mathbf{B}$.

In cases where only one of the regression parameters specified in the Level 1 model is treated as varying across groups, the conditional posterior distribution for the Level 2 scale parameter τ based on the use of an inverse chi-square prior is as follows (cf. Seltzer, 1993):

$$1/\tau \sim \text{Gamma}\left[a = [(J + \nu_2)/2], b = \left[2 \Big/ \left(S_2 + \sum_{j=1}^{J} q_j U_j^2\right)\right]\right]. \quad (30)$$

Step 4. $p(\boldsymbol{\gamma}|\mathbf{y}, \boldsymbol{\gamma}_+, \mathbf{q}, \sigma^2, \boldsymbol{\beta}, \mathbf{T})$:

$$\boldsymbol{\gamma} \sim N_K(\tilde{\boldsymbol{\gamma}}, \tilde{\mathbf{D}}), \quad (31)$$

where

$$\tilde{\boldsymbol{\gamma}} = \left[\sum_{j=1}^{J} q_j \mathbf{W}_j' \mathbf{T}^{-1} \mathbf{W}_j\right]^{-1} \sum_{j=1}^{J} q_j \mathbf{W}_j' \mathbf{T}^{-1} \boldsymbol{\beta}_j,$$

and

$$\tilde{\mathbf{D}} = \left[\sum_{j=1}^{J} q_j \mathbf{W}_j' \mathbf{T}^{-1} \mathbf{W}_j\right]^{-1}.$$

In contrast to Step 4 in the MVN algorithm (Equation 21), it is seen that those Level 2 units with small values of q_j receive proportionately less weight in the calculation of $\tilde{\boldsymbol{\gamma}}$ and $\tilde{\mathbf{D}}$.

Step 5. $p(\boldsymbol{\gamma}_+|\mathbf{y}, \mathbf{q}, \sigma^2, \boldsymbol{\beta}, \mathbf{T}, \boldsymbol{\gamma})$: This distribution is identical to Equation 23 in the MVN formulation.

Step 6. $p(\mathbf{q}|\mathbf{y}, \sigma^2, \boldsymbol{\beta}, \mathbf{T}, \boldsymbol{\gamma}, \boldsymbol{\gamma}_+)$:

$$q_j \sim \text{Gamma}[a = [(\nu + P)/2], b = [2/(\mathbf{U}_j' \mathbf{T}^{-1} \mathbf{U}_j + \nu)]], \quad (32)$$

for $j = 1, \ldots, J$. The quantity $\mathbf{U}'_j \mathbf{T}^{-1} \mathbf{U}_j$ is the squared distance of the vector $\boldsymbol{\beta}_j$ from the prior mean $\mathbf{W}_j \boldsymbol{\gamma}$. Consequently, when one or more elements of

$\boldsymbol{\beta}_j$ are far from the corresponding elements of $\mathbf{W}_j\boldsymbol{\gamma}$, the scale parameter b will take on a small value (cf. Seltzer, 1993, p. 219). This, in turn, results in a small expected value for the distribution of q_j, where $E(q_j) = ab$.

As in implementing the algorithm for the MVN Level 2 setting, we cycle through the steps of the algorithm, sampling once from each conditional posterior given the current values for the other parameters in the model.

EB and Fully Bayesian Comparisons via an Illustrative Example

We now use the above algorithms as well as standard EB procedures in an analysis of the data from an evaluation of Transition Mathematics (TM), an innovative prealgebra curriculum developed by the University of Chicago School Mathematics Project (1986). A key goal of this section is to explicate when and why it becomes important to employ a fully Bayesian approach.

In order to study the effectiveness of the TM curriculum, a field experiment was conducted during the 1985–1986 school year. The study's sample consisted of approximately 600 students nested within 20 matched pairs of classrooms located in school districts throughout the United States (while several districts contained 2 matched pairs, in most cases there was 1 matched pair per district). Each matched pair—referred to as a site in this article—consisted of two classrooms of students of comparable mathematics ability; matching was based on pretests administered at the start of the school year, as well as information provided by teachers and district mathematics coordinators. The teacher of one class used the TM text with his or her class, while the teacher of the second class used the materials already in place at that school. Thus, the matched pairs might be viewed as blocks. Note that the teachers participating in the study had considerable teaching experience (see University of Chicago School Mathematics Project, 1986, for details).

Posttests were administered at the end of the 1985–1986 school year. In our analysis, we focus on geometry readiness, which was measured by a student's total score on a 19-item test adapted from an instrument employed in a large-scale study of geometry achievement among U.S. secondary school students conducted by researchers at Ohio State University (University of Chicago School Mathematics Project, 1986).

We take into account the nested structure of the data by posing the following Level 1 model. For each of J sites ($j = 1, \ldots, 20$) we have

$$y_{ij} = \beta_{j0} + \beta_{j1}(TRT_{ij} - \overline{TRT_j}) + e_{ij}, \tag{33}$$

where y_{ij} is the geometry readiness score for student i at site j, and $TRT_{ij} = 1$ if student i was a member of the TM class at site j (0 otherwise). As in Raffe's (1991) analysis of the data from a multisite educational evaluation conducted in Great Britain, TRT_{ij} is centered around the site mean value for the treatment indicator variable ($\overline{TRT_j}$). By virtue of this, β_{j0} represents the mean geometry readiness score for site j (i.e., μ_j), and β_{j1} is the TM/Compari-

son class contrast for site j (i.e., $\mu_{j(TM)} - \mu_{j(C)}$). The OLS estimates for these parameters reduce to $\hat{\beta}_{j0} = \bar{Y}_j$ and $\hat{\beta}_{j1} = (\bar{Y}_{j(TM)} - \bar{Y}_{j(C)})$, respectively. The e_{ij} are errors assumed normally distributed with mean 0 and variance σ^2.

A modest increase in precision (i.e., a reduction in σ^2) could be achieved by using student pretest scores as a covariate in the Level 1 model (i.e., $\beta_{j2}PRETEST_{ij}$). Appendix A discusses the use of the algorithms outlined above in fitting models in which the effects of pretest are treated as constant (fixed) across groups. To help clarify the ways and extent to which answers can change as we move from EB to fully Bayesian approaches, we use the simpler Level 1 model specified in Equation 33 in our analyses. While the differences between results presented below and those based on a model that treats $PRETEST_{ij}$ as a fixed effect are minor, results based on the latter model would also need to be considered in formulating any final judgments concerning the merits of TM.

As can be seen (Table 1), the OLS estimates of the site means and contrasts ($\hat{\beta}_{j0}$, $\hat{\beta}_{j1}$) vary substantially across sites. One possible explanation for the variability in contrasts centers on the key role that reading plays in the TM curriculum; it is suspected that TM may be most effective when teachers discuss the reading passages in the text with their students on a daily basis. Information regarding this aspect of implementation is used as a predictor in the Level 2 model. In addition, we use site pretest means to model variability in β_{j0}. Thus,

$$\begin{aligned} \beta_{j0} &= \gamma_{00} + \gamma_{01} PREMN_j + U_{j0}, \\ \beta_{j1} &= \gamma_{10} + \gamma_{11} IMPLMNT_j + U_{j1}, \end{aligned} \tag{34}$$

where $PREMN_j$ is the mean for site j on a general mathematics pretest and $IMPLMNT_j$ is an indicator variable that takes on a value of 1 if the TM teacher at site j discussed reading in class on a daily basis (0 otherwise). By virtue of the dummy coding for $IMPLMNT_j$, γ_{10} represents the effect of TM at low-implementation sites ($IMPLMNT_j = 0$), and γ_{11} captures the expected increase in effectiveness given a high level of implementation ($IMPLMNT_j = 1$). Finally, U_{j0} and U_{j1} represent deviations in site means and contrasts from prior means conditional on $PREMN_j$ and $IMPLMNT_j$.

Assuming that $\boldsymbol{\beta}_j = (\beta_{j0}\ \beta_{j1})'$ is MVN distributed, we have $\boldsymbol{\beta}_j|\boldsymbol{\gamma}, \mathbf{T} \sim N_2(\mathbf{W}_j\boldsymbol{\gamma}, \mathbf{T})$, where

$$\boldsymbol{\gamma} = (\gamma_{00} \quad \gamma_{01} \quad \gamma_{10} \quad \gamma_{11})', \tag{35}$$

$$\mathbf{W}_j = \begin{pmatrix} 1 & PREMN_j & 0 & 0 \\ 0 & 0 & 1 & IMPLMNT_j \end{pmatrix}, \tag{36}$$

TABLE 1
Summary of the Transition Mathematics data

Site number	Size (n_j)	Geometry readiness outcomes: Site mean, OLS $\hat{\beta}_{j0}$	Site mean, OLS $SE(\hat{\beta}_{j0})$	TM/Comparison class contrast, OLS $\hat{\beta}_{j1}$	TM/Comparison class contrast, OLS $SE(\hat{\beta}_{j1})$	TM/Comparison class contrast, EB β^*_{j1ml}	TM/Comparison class contrast, EB $SE(\beta^*_{j1ml})$
		High-implementation sites					
8	23	6.87	0.62	−2.52	1.25	−0.24	0.91
18	28	6.89	0.55	0.78	1.12	1.42	0.88
12	32	14.52	0.53	0.79	1.06	1.18	0.84
7	37	13.97	0.39	1.27	0.79	1.59	0.80
6	36	11.44	0.55	1.68	1.12	1.76	0.81
2	28	6.54	0.61	2.16	0.89	2.18	0.87
10	20	8.45	0.75	3.63	1.52	2.63	0.96
14	26	10.15	0.56	4.09	1.27	3.11	0.94
9	43	7.12	0.43	4.42	0.85	3.57	0.75
20	17	7.12	0.82	4.56	1.72	3.18	1.00
		Low-implementation sites					
11	28	6.68	0.56	−2.21	1.13	−1.17	0.87
16	34	11.56	0.63	−1.39	1.27	−0.89	0.82
4	44	7.86	0.49	−0.24	0.99	−0.01	0.75
1	31	13.52	0.48	−0.23	0.96	−0.16	0.84
5	17	8.47	0.61	0.29	1.22	0.17	0.98
3	35	5.17	0.50	0.38	0.75	0.28	0.81
15	24	10.88	0.68	0.94	1.37	0.59	0.91
13	31	10.87	0.76	1.15	1.52	0.83	0.84
17	35	8.54	0.55	1.65	1.09	1.19	0.81
19	18	5.22	0.52	2.47	1.06	1.40	0.98

Note. The sites have been grouped by the level of implementation of the Transition Mathematics (TM) curriculum in TM classrooms; within each grouping, the sites have been sorted based on the OLS estimates of the TM/Comparison class contrasts ($\hat{\beta}_{j1}$). Note that β^*_{j1ml} and $SE(\beta^*_{j1ml})$ are, respectively, equal to the mean and standard deviation of the conditional posterior $p(\beta_{j1}|\mathbf{y}, \mathbf{T} = \mathbf{T}_{ml}, \sigma^2 = \sigma^2_{ml})$.

and where the elements of $\mathbf{T}$ are as follows:

$$\begin{aligned} T_{11} &= \text{Var}(\beta_{j0} \mid PREMN_j), \\ T_{22} &= \text{Var}(\beta_{j1} \mid IMPLMNT_j), \\ T_{21} &= \text{Cov}(\beta_{j0}, \beta_{j1} \mid \mathbf{W}_j). \end{aligned} \tag{37}$$

The plan for this section of the article is as follows. We first discuss several points concerning the implementation of the above Gibbs sampling

algorithms. Next, we focus on results for elements of $\boldsymbol{\gamma}$, comparing answers based on conditional posteriors evaluated at the ML estimates of the variance components, marginal posteriors based on MVN Level 2 distributional assumptions, and marginal posteriors based on MVT Level 2 distributional assumptions. We then proceed in a similar fashion in examining results for random regression parameters.

In utilizing the MVN and MVT Gibbs sampling algorithms outlined above, marginal posteriors of interest were calculated under assumptions of uniform (U) and inverse Wishart (IW) priors for the Level 2 variance components. Due to the fairly large number of observations in this sample ($N = 587$), our results are insensitive to choice of prior for σ^2. We therefore present results only for $p(\sigma^2) \propto k$. A uniform prior is also assumed for $\boldsymbol{\gamma}$. (For details on sampling from the conditional posterior distribution of $\mathbf{T}^{-1}$, see Appendix B.)

In order to monitor convergence of the Gibbs sampler, it is important to run multiple sequences or chains using a range of different starting values (Gelman & Rubin, 1992), as opposed to running one long sequence. Using a multiple-sequence approach similar to that outlined in Seltzer (1993), convergence appeared to occur within 2,000 iterations.[3]

Due to dependencies among deviates generated by MCMC methods, large sets of deviates are often needed in order to realize high degrees of accuracy in simulating marginal posteriors of interest (see Tanner & Wong, 1987). Because of our interest in highly precise comparisons of EB and fully Bayesian results, all histograms, intervals, and measures of location and spread for marginal posteriors reported below are based on samples of 40,000 deviates. (Note that one strategy in using the Gibbs sampler is to apply density estimation techniques to smaller samples of deviates; see, for example, Zeger & Karim, 1991.)

Fixed Effects

Employing an EB approach, we used the EM algorithm to find the values of the variance components that maximize $l(\mathbf{T}, \sigma^2|\mathbf{y})$, yielding $T_{11\text{ml}} = 1.22$, $T_{22\text{ml}} = 1.76$, $T_{21\text{ml}} = -.26$, and $\sigma^2{}_{\text{ml}} = 9.20$. (The HLM computer program [Bryk, Raudenbush, Seltzer, & Congdon, 1988] was used to accomplish this task.) Examining results based on Equation 7, we find that the mean of the conditional posterior distribution for γ_{10}—the expected effect of TM in low-implementation sites—is approximately a quarter of a point and that the resulting 95% interval comfortably includes 0 (Table 2). However, the mean of the conditional posterior distribution of γ_{11}—the expected increase in the effectiveness of TM when implementation is high—is approximately one and three quarters points, and the 95% interval that we obtain excludes a value of 0. Using the fact that the conditional posterior for γ_{11} is normal, we find that $p(\gamma_{11} < 0|\mathbf{y}, \mathbf{T} = \mathbf{T}_{\text{ml}}, \sigma^2 = \sigma^2{}_{\text{ml}}) = .011$.

TABLE 2

Conditional and marginal posterior distributions of the fixed effects under MVN and MVT_4 Level 2 assumptions. Marginal posteriors were calculated using inverse Wishart (IW) and uniform (U) priors for the Level 2 variance components.

Fixed effect	Mean	*SD*	95% interval
	Constant (γ_{00})		
MVN			
$p(\gamma_{00} \mid \mathbf{y}, \mathbf{T} = \mathbf{T}_{ml}, \sigma^2 = \sigma^2_{ml})$	−2.74	1.35	(−5.38, −0.10)
$p(\gamma_{00} \mid \mathbf{y})_{IW}$	−2.77	1.49	(−5.69, 0.15)
$p(\gamma_{00} \mid \mathbf{y})_{U}$	−2.76	1.62	(−5.97, 0.47)
MVT_4			
$p(\gamma_{00} \mid \mathbf{y})_{IW}$	−2.61	1.48	(−5.60, 0.24)
$p(\gamma_{00} \mid \mathbf{y})_{U}$	−2.59	1.61	(−5.85, 0.52)
	Effect of pretest on site means (γ_{01})		
MVN			
$p(\gamma_{10} \mid \mathbf{y}, \mathbf{T} = \mathbf{T}_{ml}, \sigma^2 = \sigma^2_{ml})$	0.62	0.07	(0.48, 0.75)
$p(\gamma_{10} \mid \mathbf{y})_{IW}$	0.62	0.08	(0.47, 0.77)
$p(\gamma_{10} \mid \mathbf{y})_{U}$	0.62	0.08	(0.45, 0.78)
MVT_4			
$p(\gamma_{00} \mid \mathbf{y})_{IW}$	0.61	0.07	(0.46, 0.76)
$p(\gamma_{10} \mid \mathbf{y})_{U}$	0.61	0.08	(0.45, 0.77)
	Expected TM effect given low implementation (γ_{10})		
MVN			
$p(\gamma_{10} \mid \mathbf{y}, \mathbf{T} = \mathbf{T}_{ml}, \sigma^2 = \sigma^2_{ml})$	0.24	0.56	(−0.84, 1.33)
$p(\gamma_{10} \mid \mathbf{y})_{IW}$	0.24	0.62	(−0.97, 1.46)
$p(\gamma_{10} \mid \mathbf{y})_{U}$	0.25	0.67	(−1.06, 1.60)
MVT_4			
$p(\gamma_{10} \mid \mathbf{y})_{IW}$	0.25	0.58	(−0.89, 1.41)
$p(\gamma_{10} \mid \mathbf{y})_{U}$	0.26	0.63	(−0.98, 1.53)
	Increment in TM effect given high implementation (γ_{11})		
MVN			
$p(\gamma_{11} \mid \mathbf{y}, \mathbf{T} = \mathbf{T}_{ml}, \sigma^2 = \sigma^2_{ml})$	1.78	0.78	(0.24, 3.31)
$p(\gamma_{11} \mid \mathbf{y})_{IW}$	1.78	0.86	(0.07, 3.48)
$p(\gamma_{11} \mid \mathbf{y})_{U}$	1.77	0.93	(−0.06, 3.62)
MVT_4			
$p(\gamma_{11} \mid \mathbf{y})_{IW}$	1.87	0.84	(0.21, 3.55)
$p(\gamma_{11} \mid \mathbf{y})_{U}$	1.87	0.91	(0.07, 3.65)

Taking Into Account Uncertainty in the Variance Components via the Gibbs Sampler

We first present results based on uniform priors for $\mathbf{T}$ and σ^2, and begin by focusing attention on the marginal posterior distributions of the variance

components (see Table 3 and Figure 1). Under assumptions of uniform priors, these distributions correspond to the marginal likelihood functions of the variance components, for example, $p(T_{22}|\mathbf{y}) \propto l(T_{22}|\mathbf{y})$. Examining $p(T_{11}|\mathbf{y})$ and $p(T_{22}|\mathbf{y})$ enables us to see the crux of the problem that Rubin (1981) draws attention to in his fully Bayesian analysis of SAT coaching effects: Likelihood functions for Level 2 variances will tend to be highly positively skewed when J is small such that most of the mass of the likelihood will lie above the ML estimate. Hence, treating ML estimates for the Level 2 variances as known values results in an "underpropagation" of uncertainty in constructing intervals for parameters of interest (e.g., fixed effects) (Draper, 1995) and can yield point estimates for parameters of interest that are poor summaries of the data (Rubin, 1981).

Under the assumption of uniform priors for the variance components, we find that the standard deviations of the marginal posteriors of the fixed effects are appreciably larger than the standard deviations based on the conditional posteriors (Table 2). In addition, we see that the resulting 95% interval for γ_{11} includes a value of 0. However, based on the empirical distribution of the deviates generated for γ_{11}, we find that $p(\gamma_{11} < 0|\mathbf{y})_U = .028$; hence, it would be unwise to dismiss the notion of increased effectiveness of TM in high-implementation sites.

To grasp how taking into account uncertainty concerning the variance components results in increases in dispersion in the posterior distributions of the elements of $\boldsymbol{\gamma}$, it is helpful to consider that calculating $p(\boldsymbol{\gamma}|\mathbf{y})$ is equivalent to averaging $p(\boldsymbol{\gamma}|\mathbf{y}, \mathbf{T}, \sigma^2)$ over $p(\mathbf{T}, \sigma^2|\mathbf{y})$. From Equation 4, it is clear that different values for the variance components that we might condition on result in conditional posterior distributions for $\boldsymbol{\gamma}$ that differ in terms of dispersion (e.g., $\mathbf{D}^*$) and possibly location, as well. In averaging over $p(\mathbf{T}, \sigma^2|\mathbf{y})$, the posterior probabilities for different possible values for the variance components serve as weights for the corresponding conditional posterior

TABLE 3
Marginal posterior distributions of the variance components under MVN Level 2 assumptions and based on uniform priors for the variance components

			Quantiles				
Distribution	Mode	Mean	.025	.25	.5	.75	.975
$p(\sigma^2 \mid \mathbf{y})$	9.20	9.26	8.21	8.88	9.24	9.63	10.43
$p(T_{11} \mid \mathbf{y})$	1.45	2.00	0.74	1.30	1.77	2.41	4.66
$p(T_{22} \mid \mathbf{y})$	2.25	3.33	0.72	1.89	2.86	4.22	8.70
$p(T_{21} \mid \mathbf{y})$	−0.35	−0.39	−2.54	−0.88	−0.33	0.16	1.41

Note. These are the marginals that would be obtained by integrating over the joint likelihood for the variance components, i.e., $l(\mathbf{T}, \sigma^2|\mathbf{y})$. The marginal modes for T_{11} and T_{22} are larger than the joint modal values ($T_{11\text{ml}}$ and $T_{22\text{ml}}$) due to the asymmetry in $l(\mathbf{T}, \sigma^2|\mathbf{y})$. In cases where the joint likelihood is symmetric (i.e., when J is large), the joint modes and marginal modes for the variance components would be identical.

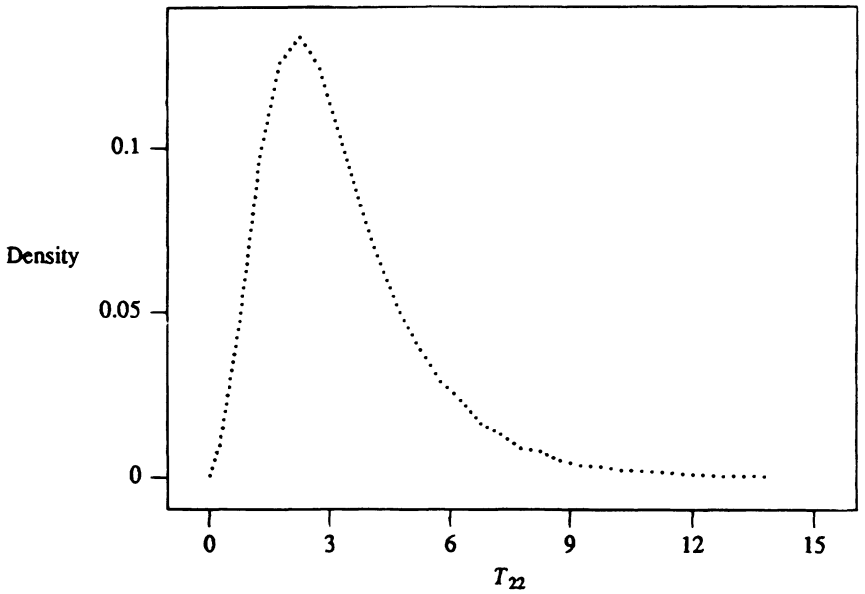

FIGURE 1. *Histogram of the marginal posterior distribution of* T_{22} *under MVN Level 2 assumptions and based on uniform priors for the variance components*

distributions for $\boldsymbol{\gamma}$ that could be formed. Thus, given the asymmetry of the posterior distribution of the variance components (i.e., the positive skew connected with T_{11} and T_{22}), considerable weight will be attached to conditional posterior distributions of $\boldsymbol{\gamma}$ based on values of $T_{11} > T_{11\text{ml}}$ ($T_{11\text{ml}} = 1.22$) and $T_{22} > T_{22\text{ml}}$ ($T_{22\text{ml}} = 1.76$). (For a discussion of this conceptualization in the context of the one-sample problem—that is, viewing $p(\mu|\mathbf{y})$ as a weighted average of $p(\mu|\sigma^2, \mathbf{y})$ over $p(\sigma^2|\mathbf{y})$—see Box and Tiao, 1973, chap. 2.)

As noted earlier, a potential drawback of the uniform prior is that even for arbitrarily large values for the Level 2 variances, prior probabilities do not diminish. Hence, the credibility intervals that we obtain may be somewhat conservative. Thus, as an alternative, we now assume that $\mathbf{T}$ is a priori IW distributed with $\nu_2 = 3$ degrees of freedom and scale matrix $\mathbf{S}_2$ with diagonal elements $S_{11} = 5.8$ and $S_{22} = 9.0$ and off-diagonal elements $S_{12} = S_{21} = 0$. With $\nu_2 = 3$, the information provided by this prior is weak relative to the information contributed by the likelihood. Using results presented in Zellner (1971), the values of the elements of $\mathbf{S}_2$ were chosen so that the prior slightly favors values of T_{11} and T_{22} that are approximately equal to the modes of $l(T_{11}|\mathbf{y})$ and $l(T_{22}|\mathbf{y})$, respectively.[4] While giving some support to the information provided by the data, this prior, unlike the uniform prior, downweights arbitrarily large values for T_{11} and T_{22}.

Due to the downweighting a priori of extreme values for the Level 2 variances, we see that the IW prior results in a decrease in dispersion for the

marginal posterior distributions of the fixed effects (Table 2). It is also seen that the lower boundary of the 95% interval for γ_{11} is slightly larger than 0.

While this example helps to illustrate the conservative nature of the uniform prior, the answers that we obtain under the two priors do not differ greatly from a substantive standpoint. In particular, based on the empirical distributions for γ_{11}, $p(\gamma_{11} < 0|\mathbf{y}) = .022$ under the IW prior versus a value of .028 under the uniform prior.[5] However, as J decreases and, in turn, the amount of information provided by the data concerning the Level 2 variance components decreases—that is, as the likelihood for the variance components becomes increasingly flat and disperse—one's results become increasingly sensitive to choice of priors (e.g., U, IW) for $\mathbf{T}$.

In anticipating the extent to which answers concerning fixed effects of interest might differ based on EB and fully Bayesian approaches, it is helpful to consider again the one-sample problem, that is, $y_i \sim N(\mu, \sigma^2)$. We know that in treating the sample variance s^2 as the known value for σ^2 and using critical values based on the z distribution to construct intervals for μ, understatements of uncertainty in the intervals that we obtain tend to be inconsequential when, for example, $n - 1 > 30$. However, understatements of uncertainty become more consequential as $n - 1$ decreases. In particular, in settings where $n - 1$ is quite small (e.g., < 10), intervals based on t critical values can potentially lead to conclusions that differ substantially from conclusions reached using z critical values.

Whereas the number of degrees of freedom available for estimating σ^2 plays a crucial role in the one-sample problem, the number of Level 2 units in a sample (J) and the number of fixed effects that are being estimated (K) play a key role in drawing inferences concerning fixed effects in the HM (Bryk & Raudenbush, 1992). With $J - K > 30$, understatements of uncertainty due to treating the ML estimates of the variance components as known values (Equation 7) are likely to be quite minor. But as $J - K$ decreases, propagation of uncertainty in the variance components via a fully Bayesian analysis will begin to render less plausible the conclusions that we might base initially on Equation 7. As in the one-sample problem, it is in settings where $J - K$ is very small (e.g., 10 or less) that taking into account uncertainty in the variance components can potentially result in conclusions concerning fixed effects of interest that differ markedly from those based on the use of an EB approach (Seltzer, 1991).

Note that in the case of simple, balanced hierarchical designs (see, for example, Kirk, 1982), one can, in the frequentist framework, construct intervals of desired levels of confidence by employing standard errors based on Equation 7 and critical values based on a t distribution with $J - K$ degrees of freedom (see Raudenbush, 1992). While the kinds of hierarchically structured data sets that we typically encounter in educational research are unbalanced, Bryk and Raudenbush (1992, chap. 9) point out that, provided one's data are not too unbalanced, using critical values based on the family of t

distributions should in many instances provide a reasonable ad hoc approach to adjusting for uncertainty in the variance components when drawing inferences concerning $\boldsymbol{\gamma}$ in small-sample settings. When it is not possible to carry out fully Bayesian analyses due to, for example, unavailability of software, this ad hoc approach would be preferable to a standard EB strategy that entails the use of z critical values; such an approach can, however, still result in underestimates of uncertainty (see Bryk & Raudenbush, 1992, chap. 9).

Studying the Sensitivity of Results to Extreme Level 2 Units

Using diagnostics suggested by Waternaux, Laird, and Ware (1989) for checking the adequacy of Level 2 models, a plot of EB residuals (i.e., $\beta^*_{j1\text{ml}}$ minus the estimated prior mean $[\gamma^*_{10\text{ml}} + \gamma^*_{11\text{ml}}\, IMPLMNT_j]$) revealed a large negative residual for Site 8, a high-implementation site (see Figure 2). The Gibbs sampler was used to conduct a sensitivity analysis in which the marginal posterior distributions of the fixed effects were recalculated under heavy-tailed (MVT_4) distributional assumptions at Level 2, that is, $\boldsymbol{\beta}_j|\boldsymbol{\gamma}, \mathbf{T} \sim t_P(\mathbf{W}_j\boldsymbol{\gamma}, \mathbf{T}, \nu = 4)$ (see Lange, Little, & Taylor, 1989, pp. 882–883). Through a series of reanalyses with ν fixed at several different values (e.g., 20, 7, 4, 2), one can examine the sensitivity of one's results to varying degrees of heavy-tailedness, as in Seltzer (1993). For illustrative purposes, we have chosen to reanalyze the data under the assumption of rather heavy tails at Level 2 (i.e., $\nu = 4$). We present results based on a uniform prior for the Level 2 variance components and on an IW prior with $\nu_2 = 3$ and scale matrix values of $S_{11} = 4.4$, $S_{22} = 5$, and $S_{12} = S_{21} = 0$ (see Note 6).

Comparing the $MVT_{4(U)}$ and $MVT_{4(IW)}$ results (Table 2), we see that both analyses yield a similar increase in the posterior mean for γ_{11}, but as in the MVN analyses, the use of the IW prior for the Level 2 variance components results in a smaller posterior standard deviation. Interestingly, the lower boundary of the 95% interval for γ_{11} produced by the $MVT_{4(IW)}$ analysis is

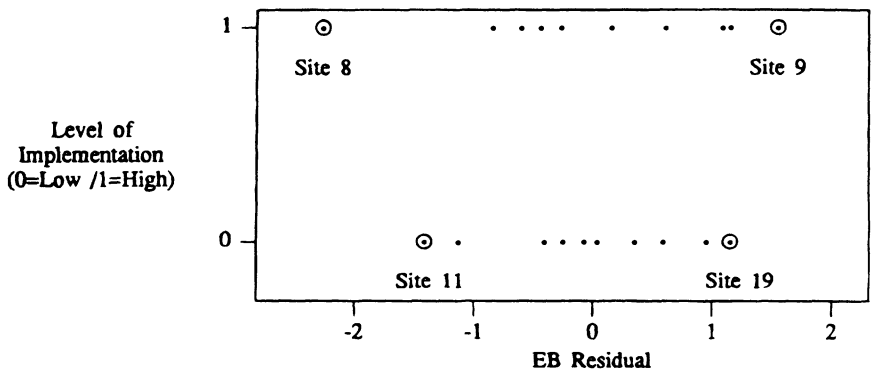

FIGURE 2. *Empirical Bayes residual TM/Comparison class contrasts displayed by level of implementation*

nearly as large as the lower boundary based on the initial analysis in which we condition on the ML estimates of the variance components, while the upper boundary is appreciably larger.

Just as in the case of the normal/gamma univariate t formulation (see, e.g., Seltzer, 1993), outlying Level 2 units in the multivariate t formulation are downweighted via small values of q_j. In this regard, the median of $p(q_8|\mathbf{y})$ in our analysis is equal to .49, while the medians of $p(q_j|\mathbf{y})$ for the remaining 19 sites range between values of .75 and 1.0. (The quantiles of $p(q_j|\mathbf{y})$ for Sites 8 and 11 are displayed in Table 4; we focus on drawing inferences concerning the TM/Comparison class contrasts (β_{j1}) for these sites in the next subsection of the article.)

Note that the degree of sensitivity of fixed effects to outlying Level 2 units will depend in part on the number of outliers in a sample, how extreme they are, and the extent to which they are also extreme on $\mathbf{W}_j$, the set of Level 2 predictors (i.e., the degree to which they leverage the Level 2 fit). For a detailed discussion of the use of the q_j in identifying outlying Level 2 units, see Seltzer (1993).

Random Regression Parameters

Propagating Uncertainty in the Variance Components

To explicate the kinds of changes that can occur in location and dispersion for posterior distributions of random regression parameters upon taking into account uncertainty in the variance components, we focus on the TM/Comparison class contrast for Site 8, an outlying site (Figure 2) which has the smallest OLS contrast among the high-implementation sites ($\hat{\beta}_{(8)1} = -2.52$), and for Site 11, which has the smallest OLS contrast among the low-implementation sites ($\hat{\beta}_{(11)1} = -2.21$) (Table 1).

Substituting $\mathbf{T}_{ml}$ and σ^2_{ml} into Equation 5 yields the EB estimates $\beta^*_{(8)1ml} = -.24$ and $\beta^*_{(11)1ml} = -1.17$. The reason why the EB estimates for these sites differ substantially becomes clear when we consider that the prior mean for within-site contrasts based on the Level 2 model is $\gamma_{10} + \gamma_{11}IMPLMNT_j$

TABLE 4
Posterior distributions of q_j *and* T_{22}/q_j *for Sites 8 and 11*

Posterior distribution	Mean	Quantiles				
		.025	.25	.5	.75	.975
		Site 8				
$p\ (q_8 \mid \mathbf{y})$	0.63	0.07	0.27	0.49	0.84	2.00
$p\ (T_{22} / q_8 \mid \mathbf{y})$	7.14	0.59	2.36	4.47	8.46	29.99
		Site 11				
$p\ (q_{11} \mid \mathbf{y})$	1.07	0.18	0.56	0.92	1.41	2.83
$p\ (T_{22} / q_{11} \mid \mathbf{y})$	3.78	0.37	1.27	2.36	4.38	15.48

Note. Results are based on uniform priors for the Level 1 and Level 2 variance components.

(see Equation 34). Substituting the EB estimates of the fixed effects into Equation 34, we see that the OLS contrast for Site 8 (*IMPLMNT* = 1) is shrunk toward an estimated prior mean of .24 + 1.78 = 2.02, whereas the OLS contrast for Site 11 (*IMPLMNT* = 0) is shrunk toward an estimated prior mean of .24.

We next examine the marginal posterior distributions for $\beta_{(8)1}$ and $\beta_{(11)1}$ under the assumption of uniform priors for the variance components.[7] First, it can be seen that the posterior means have to some extent been pulled back toward the OLS estimates of the TM/Comparison class contrasts (see Table 5). In particular, we see that the posterior mean for Site 8 has shifted from a value of −.24 to a value of −.58 (Figure 3). The reason for these shifts becomes clear in view of the following. The magnitude of T_{22} (i.e., the prior variance for within-site contrasts) directly affects how the data for a particular site ($\hat{\beta}_{j1}$) and the prior mean (i.e., $\gamma_{10} + \gamma_{11}IMPLMNT_j$) are weighted in forming a composite estimate of β_{j1}. Given the structure of the conditional posterior mean $\boldsymbol{\beta}_j^*$ (Equation 5), it is clear that as the values of T_{22} that we condition on increase, the amount of weight placed on $\hat{\beta}_{j1}$ increases. Since $E(\boldsymbol{\beta}_j|\mathbf{y})$, the marginal posterior mean of $\boldsymbol{\beta}_j$, can be viewed as a weighted average of $\boldsymbol{\beta}_j^*$ over $p(\mathbf{T}, \sigma^2|\mathbf{y})$, and noting that more than 75% of the mass

TABLE 5

TM/Comparison class contrasts for Sites 8 and 11 under MVN and MVT_4 Level 2 assumptions. Marginal posteriors were calculated using weak inverse Wishart (IW) and uniform (U) priors for the Level 2 variance components.

TM/Comparison class contrast	Mean	*SD*	95% interval
	Contrast for Site 8		
MVN			
$p(\beta_{(8)1} \mid \mathbf{y}, \mathbf{T} = \mathbf{T}_{ml}, \sigma^2 = \sigma^2_{ml})$	−0.24	0.91	(−2.02, 1.54)
$p(\beta_{(8)1} \mid \mathbf{y})_{IW}$	−0.49	1.11	(−2.73, 1.61)
$p(\beta_{(8)1} \mid \mathbf{y})_{U}$	−0.58	1.22	(−3.04, 1.74)
MVT_4			
$p(\beta_{(8)1} \mid \mathbf{y})_{IW}$	−0.85	1.30	(−3.48, 1.57)
$p(\beta_{(8)1} \mid \mathbf{y})_{U}$	−0.94	1.35	(−3.62, 1.65)
	Contrast for Site 11		
MVN			
$p(\beta_{(11)1} \mid \mathbf{y}, \mathbf{T} = \mathbf{T}zm_l, \sigma^2 = \sigma^2_{ml})$	−1.17	0.87	(−2.88, 0.54)
$p(\beta_{(11)1} \mid \mathbf{y})_{IW}$	−1.26	0.97	(−3.21, 0.59)
$p(\beta_{(11)1} \mid \mathbf{y})_{U}$	−1.33	1.01	(−3.36, 0.58)
MVT_4			
$p(\beta_{(11)1} \mid \mathbf{y})_{IW}$	−1.18	0.98	(−3.18, 0.68)
$p(\beta_{(11)1} \mid \mathbf{y})_{U}$	−1.22	1.02	(−3.37, 0.64)

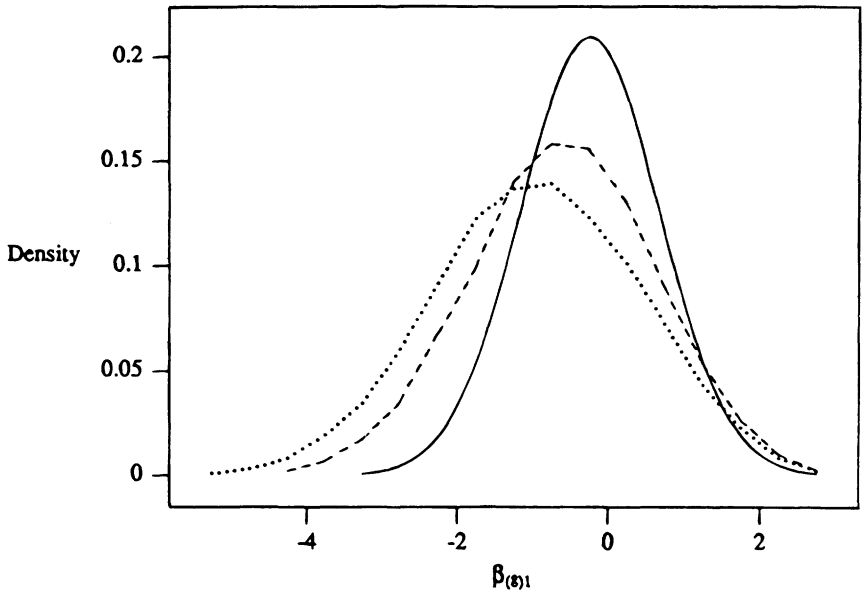

FIGURE 3. *Histograms of the posterior distributions of the TM/Comparison class contrast for Site 8 (solid line = conditional posterior assuming an MVN prior; dashed line = marginal posterior assuming an MVN prior; dotted line = marginal posterior assuming an MVT_4 prior). Marginal posteriors were calculated based on uniform priors for the variance components.*

of $p(T_{22}|\mathbf{y})$ lies above $T_{22\text{ml}}$ (see Figure 1 and Table 3), we find that considerably more weight is placed on the data in calculating $E(\beta_{j1}|\mathbf{y})$ than in computing $\beta^*_{j1\text{ml}}$.

Using the IW prior for the Level 2 variance components results in slightly less of a decrease in the posterior means for Sites 8 and 11. This is due to the fact that the IW prior results in a posterior distribution for the variance components that places somewhat less probability on extreme values of T_{22}.

For those sites where $\hat{\beta}_{j1}$ is approximately equal to the prior mean, β^*_{j1} remains essentially unchanged over all possible values of T_{22}. For such sites, $E(\beta_{j1}|\mathbf{y}) \approx \beta^*_{j1\text{ml}}$ (see Table 1). In many applications, however, substantive interest often centers on cases where the OLS estimate for a Level 1 parameter is substantially larger or smaller than the corresponding prior mean. When J is small, it is precisely in such cases that taking into account uncertainty in the variance components can result in substantial shifts in the posterior means of random regression parameters as depicted above.

We now focus on changes in dispersion of posterior distributions for TM/Comparison contrasts resulting from taking into account uncertainty in the variance components. We first consider results under uniform priors for the variance components. In Table 5, it can be seen that the marginal posterior standard deviations of the within-site contrasts are substantially larger than

the standard deviations of the corresponding conditional posterior distributions based on the ML estimates of the variance components.

While the standard deviations of the conditional posteriors for Sites 8 and 11 are nearly identical, calculation of the marginal posterior distributions of the within-site contrasts results in a much greater increase in dispersion in the case of Site 8. This stems from the fact that a key component of $\text{Var}(\boldsymbol{\beta}_j|\mathbf{y})$ is the variance of the conditional posterior mean ($\boldsymbol{\beta}_j^*$) over $p(\mathbf{T}, \sigma^2|\mathbf{y})$ (see Rubin, 1980). Now, for sites where $\hat{\beta}_{j1}$ is approximately equal to the estimated prior mean, the composite estimator β_{j1}^* will remain virtually unchanged over possible values for T_{22} (e.g., the range of values for T_{22} spanned by $p(T_{22}|\mathbf{y})$; see Figure 1); hence, $\text{Var}(\beta_{j1}^*) \approx 0$. However, as $\hat{\beta}_{j1}$ becomes more extreme in relation to the prior mean, β_{j1}^* will change considerably as we condition on possible values for T_{22}. Thus, for Site 8, with $\hat{\beta}_{(8)1} = -2.52$ and an estimated prior mean of 2.02, $\text{Var}(\beta_{(8)1}^*)$ is substantial. An implication is that in taking into account uncertainty in the variance components, increases in posterior variance will be particularly large in the case of outlying Level 2 units.

As in the case of the fixed effects, we see somewhat less of an increase in posterior dispersion under IW priors for the Level 2 variance components. However, whereas a value of $\hat{\beta}_{(8)1} = -2.52$ is highly implausible based on the EB analysis, the 95% credibility intervals for $\beta_{(8)1}$ under uniform and IW priors comfortably include this value.

In terms of potential guidelines, intervals for random regression parameters of interest based on a fully Bayesian approach may be appreciably wider than those based on a standard approach even when $J - K$ approaches a value of 30 (Seltzer, 1991). This is particularly so for outlying Level 2 units, since in these cases $\text{Var}(\boldsymbol{\beta}_j^*)$ can be quite substantial (see Rubin, 1980, 1981).

Reanalysis Under the Assumption of Heavy-Tailed Priors

The marginal posterior distributions of the contrasts for Sites 8 and 11 were recalculated under MVT_4 distributional assumptions at Level 2. In the case of Site 11, we see that the posterior mean has been pulled back slightly toward the prior mean (Table 5). However, for Site 8, we see another substantial shift in the posterior mean toward $\hat{\beta}_{(8)1}$.

While the magnitude of T_{22} plays a crucial role in the MVN formulation in determining how the data ($\hat{\beta}_{j1}$) and the prior mean are weighted in forming a composite estimate of β_{j1}, this role is supplanted by T_{22}/q_j in the normal/gamma formulation of the *t*. Thus, as q_j decreases, the prior variance for β_{j1} increases, and more weight, in turn, will be placed on the data. Now, while the lower and upper quartiles for $p(T_{22}|\mathbf{y})$ in the MVN analysis are 1.89 and 4.22, respectively (Table 3), the lower and upper quartiles for $p(T_{22}/q_8|\mathbf{y})$ are substantially larger: 2.36 and 8.46 (Table 4). This helps us see that in calculating the marginal posterior distribution of $\beta_{(8)1}$, considerably more weight is placed on the data ($\hat{\beta}_{(8)1}$) in the MVT_4 analysis than in the MVN analysis.

The opposite, however, holds true for Site 11. Though the upper quartiles of $p(T_{22}|\mathbf{y})$ and $p(T_{22}/q_{11}|\mathbf{y})$ are similar in magnitude, the .025, .25, and .50 quantiles of $p(T_{22}/q_{11}|\mathbf{y})$ are considerably smaller; hence, somewhat more weight is placed on the prior mean in the MVT_4 setting.

We also see that the assumption of heavy tails results in a substantial increase in the standard deviation of the posterior distribution of the TM/Comparison contrast for Site 8. Since the posterior variance of β_{j1} depends on both the magnitude of the sampling variance of $\hat{\beta}_{j1}$ and the magnitude of the prior variance for β_{j1}, this increase is understandable in view of the large values spanned by $p(T_{22}/q_8|\mathbf{y})$.

Summary and Discussion

The standard approach to drawing inferences concerning fixed effects and random regression parameters of interest in the HM entails treating ML estimates of the variance components as known, true values. In this article, we have presented a general Gibbs sampling formulation that can be used to calculate marginal posterior distributions of parameters of interest in a broad range of continuous-outcome, two-level HM settings encountered in practice (see Equations 1 and 3), thus providing a means of taking into account uncertainty in the variance components. Furthermore, we have tried to clarify situations in which such an approach becomes especially important.

In using a fully Bayesian approach, it is prudent to examine the sensitivity of one's results to choice of priors for the variance components.[8] In this article, two priors for the Level 2 variance components were used: a uniform prior and a weak inverse Wishart prior with degrees of freedom and scale described above. Since prior probabilities under the uniform prior do not diminish for arbitrarily large values of the variance parameters, intervals resulting from the use of a uniform prior will tend to be more conservative. It is in settings where the number of groups in a sample is extremely small (i.e., where the information provided by the data regarding $\mathbf{T}$ is sparse) that results will potentially be most sensitive to choice of priors, and hence where sensitivity analyses become especially important.

Typically, normality is assumed at Level 2 of the HM. In this article we have illustrated the use of the Gibbs sampler in calculating marginal posterior distributions of fixed effects and random regression parameters under MVT Level 2 assumptions. Such analyses yield posterior distributions for fixed effects of interest that are resistant to outlying Level 2 units. In connection with inferences for random regression parameters, MVT analyses provide protection against overshrinkage for outlying Level 2 units.

The notion of examining the sensitivity of results to different distributional assumptions—particularly different degrees of heavy-tailedness—is a central part of the Bayesian approach (see Box & Tiao, 1973, chap. 3; Box, 1979, 1980; and Barnard, 1980). On the one hand, if posterior means and intervals for parameters of interest change substantially under different assumptions,

then this is something that needs to be brought to light in attempting to draw final conclusions. On the other hand, if our results remain essentially unchanged, this is something that helps to strengthen our conclusions. This, of course, applies to the above discussion concerning choice of priors for the Level 2 variance components.

While the use of numerical integration techniques in calculating marginal posteriors becomes problematic when joint posteriors are of high dimension, the strength of the Gibbs sampler is that it extends fairly easily to many complex settings. In addition to the kinds of models discussed in this article, the Gibbs sampler can be used in fitting HMs that consist of three or more levels and HMs with categorical Level 1 outcomes (Albert & Chib, 1993; Dellaportas & Smith, 1993; Zeger & Karim, 1991). Whereas EB estimation strategies for such models are available, the Gibbs sampler provides a viable means of taking into account uncertainty in the variance components and conducting sensitivity analyses in instances where the use of an EB approach may be problematic (e.g., when sample sizes are small). Finally, while Gibbs sampling algorithms can be implemented fairly readily using such high-level languages as XLISP-STAT, much progress has been made in developing software expressly designed for conducting fully Bayesian analyses via the Gibbs sampler (i.e., BUGS; see Thomas, Spiegelhalter, & Gilks, 1992).

Notes

[1] The mode, degrees of freedom (df), and scale parameter (S) of an inverse chi-square distribution with $df > 0$ and $S > 0$ are connected to each other in the following way: Mode $= S/(2 + df)$. Thus, if one assumes that a variance component is a priori inverse chi-square distributed with small degrees of freedom and a particular mode, it is easy to find the value of the corresponding scale parameter S.

[2] The use of the normal/gamma representation of the t and multivariate t in obtaining ML estimates of location and scale parameters in linear models is discussed in Dempster, Laird, & Rubin (1977, 1980); Rubin (1983); Little & Rubin (1987); and Lange, Little, & Taylor (1989).

[3] The MVN and MVT algorithms were written in Fortran 77 and implemented on a Hewlett Packard 9000 minicomputer. IMSL subroutines were used in order to generate values from the distributional forms (e.g., normal, gamma) found in the various steps of the algorithms, and to perform matrix operations. Completion of 1,000 iterations of the MVN algorithm required approximately 60 seconds of CPU time, while completion of 1,000 iterations of the MVT algorithm required approximately 1 minute and 15 seconds.

[4] In the current example, the dimension of $\mathbf{T}$ is $P = 2$. We now assume that $\mathbf{T}$ is a priori inverse Wishart distributed with $\nu_2 = 3$ and scale matrix $\mathbf{S}$, which consists of the diagonal elements S_{11} and S_{22} and the off-diagonal elements S_{12} and S_{21}. Based on Zellner (1971, p. 395), T_{11} is inverse chi-square distributed with $(\nu_2 - P + 1) = 2$ degrees of freedom and scale S_{11}, and T_{22} is inverse chi-square distributed with $(\nu_2 - P + 1) = 2$ degrees of freedom and scale S_{22}. Then, choosing a mode for T_{11} equivalent to the mode of $l(T_{11} | \mathbf{y})$ (i.e., 1.45; see Table 3), we solve for S_{11} based

on the simple formula provided in Note 1. We proceed in a similar fashion in choosing S_{22}.

[5] Using an IW prior that is even less informative (i.e., $\nu_2 = 2$) and that favors values of T_{11} and T_{22} slightly smaller than T_{11ml} and T_{22ml} produces results that are nearly identical to the IW results presented in Table 2.

[6] Using the logic outlined in Note 4, the values of the elements of **S** were chosen so that the prior slightly favors values of T_{11} and T_{22} that are approximately equal to the modes of $p(T_{11}|\mathbf{y})$ and $p(T_{22}|\mathbf{y})$ obtained under the assumption of a uniform prior for **T** and under MVT_4 Level 2 assumptions.

[7] Morris & Normand (1992) point out that in HM settings where the random regression parameters are scalars assumed normally distributed at Level 2, and where the sampling variances of the $\hat{\beta}_j$ are known and equivalent across units (i.e., $V_j = V$), placing a uniform prior on the Level 2 variance component leads to minimax estimators for the β_j.

[8] Rather than rerun the Gibbs sampler for each respecification of the prior, one can use the deviates from a Gibbs analysis using, for example, uniform priors, and use importance ratios (Tanner, 1993) to recalculate posterior means and variances under different specifications of the prior using the same set of deviates.

APPENDIX A
Expansion of the model used in the illustrative example

Including student pretest scores in the Level 1 model specified in Equation 33 (i.e., $\beta_{j2}PRETEST_{ij}$), and treating the pretest slope as constant across groups, we write the Level 1 model in the form used in Equation 2:

$$\mathbf{y}_j = \mathbf{X}_j\boldsymbol{\beta}_j + \mathbf{X}_{j+}\boldsymbol{\gamma}_+ + \mathbf{e}_j, \tag{A1}$$

where

$$\mathbf{X}_j = \begin{pmatrix} 1 & X_{11j} \\ 1 & X_{12j} \\ \vdots & \vdots \\ 1 & X_{1n_jj} \end{pmatrix}, \qquad \boldsymbol{\beta}_j = \begin{pmatrix} \beta_{j0} \\ \beta_{j1} \end{pmatrix},$$

with $X_{1ij} = (TRT_{ij} - \overline{TRT}_j)$, and where

$$\mathbf{X}_{j+} = \begin{pmatrix} X_{21j} \\ X_{22j} \\ \vdots \\ X_{2n_jj} \end{pmatrix}, \qquad \gamma_+ = \beta_{j2}, \tag{A2}$$

with $X_{2ij} = PRETEST_{ij}$. (Centering options for X_2 are discussed below.)

Treating β_{j0} and β_{j1} as MVN distributed, and utilizing the same Level 2 predictors specified in Equation 34, we have $\boldsymbol{\beta}_j|\boldsymbol{\gamma}, \mathbf{T} \sim N_2(\mathbf{W}_j\boldsymbol{\gamma}, \mathbf{T})$, with $\boldsymbol{\gamma}$, **W**, and **T** arrayed

as in Equations 35–37. The 5-step MVN Gibbs sampling formulation depicted in Equations 14, 17, 19, 21, and 23, with $P = 2$ and $F = 1$, can be applied directly to this model. In particular, given the data and current values for $\mathbf{T}$, $\boldsymbol{\gamma}$, γ_{+}, and σ^2, the mean of the conditional posterior distribution of the random regression parameters ($\boldsymbol{\beta}_j$) (Step 2; Equation 17) is a composite estimate based on the prior mean for site j ($\mathbf{W}_j\boldsymbol{\gamma}$) and $\hat{\boldsymbol{\beta}}_{(d)j}$, where $\hat{\boldsymbol{\beta}}_{(d)j}$ is obtained by regressing $\mathbf{d}_j = (\mathbf{y}_j - \mathbf{X}_{j+}\gamma_{+})$ on $\mathbf{X}_j$. In addition, the mean of the conditional posterior distribution of γ_{+} given the data and current values for the $\boldsymbol{\beta}_j$ (Step 5, Equation 23) is a pooled regression estimator involving the $\mathbf{d}_{j+} = (\mathbf{y}_j - \mathbf{X}_j\,\boldsymbol{\beta}_j)$ and $\mathbf{X}_{j+}$, $j = 1, \ldots, J$; the current value of σ^2 is used in calculating the variance of this conditional posterior.

The use of the MVT formulation in this setting is as depicted in Equations 14, 17, 29, 31, 23, and 32, with $\mathbf{T}$ in Equation 17 replaced by $\mathbf{T}/q_j$.

We wish to point out that the type of centering that one chooses for X_2 has implications for the interpretation of β_{j0}. Under group mean centering, the interpretation of β_{j0} remains the same as in the illustrative example. Using grand mean centering, β_{j0} represents site mean achievement adjusted for differences among sites in pretest scores, and γ_{01} in Equation 34 represents the contextual effect of the pretest variable (see Bryk & Raudenbush, 1992, pp. 115, 121–123).

Bryk and Raudenbush (1992, pp. 21–22) also discuss settings in which a Level 1 regression parameter (say, β_{j2}) might be modeled as a function of a Level 2 predictor (W_j), but with the corresponding residual Level 2 variance component ($\mathrm{Var}(\beta_{j2} | W_j)$) constrained to be 0. This can be handled easily through the inclusion of cross-level interaction terms in $\mathbf{X}_{j+}$. Thus, continuing with this example, the matrix $\mathbf{X}_{j+}$ would include columns containing values for X_{2ij} and for the cross-level interaction term $X_{2ij} \times W_j$.

APPENDIX B
Sampling from a Wishart distribution of dimension P = 2

In implementing the Gibbs sampler under MVN Level 2 assumptions, we must obtain realizations of $\mathbf{T}^{-1}$ from Wishart distributions as defined in Equation 19. Thus, in iteration i, we need to obtain a single realization $\mathbf{T}^{-1(i)}$ from a Wishart distribution with ν_w degrees of freedom and scale matrix $\mathbf{B}^{(i)}$, where the latter is a sum of squares and cross-products based on the current values (i.e., the most recently sampled values) for $\boldsymbol{\beta}_j$ ($j = 1, \ldots, J$) and $\boldsymbol{\gamma}$. In the case of the TM application where $\mathbf{T}^{-1}$ is a 2×2 matrix, we proceed as follows (see Odell & Feiveson, 1966, and Smith & Hocking, 1972, for generalizations).

(1) Draw $Z^{(i)}$ from a standard normal distribution, $K_1^{(i)}$ from a chi-square distribution with ν_w degrees of freedom (see Equation 19), and $K_2^{(i)}$ from a chi-square distribution with $\nu_w - 1$ degrees of freedom.

(2) Form the following matrix:

$$\mathbf{M}^{(i)} = \begin{pmatrix} K_1^{(i)} & Z^{(i)}K_1^{1/2(i)} \\ Z^{(i)}K_1^{1/2(i)} & Z^{2(i)} + K_2^{(i)} \end{pmatrix}. \tag{B1}$$

(3) We then obtain $\mathbf{T}^{-1(i)}$ by rescaling $\mathbf{M}^{(i)}$ using the scale matrix $\mathbf{B}^{(i)}$ (see above): $\mathbf{T}^{-1(i)} = \mathbf{C}'^{(i)}\,\mathbf{M}^{(i)}\,\mathbf{C}^{(i)}$, where $\mathbf{B}^{(i)} = \mathbf{C}'^{(i)}\,\mathbf{C}^{(i)}$.

Under MVT Level 2 assumptions, we generate $\mathbf{T}^{-1(i)}$ in a similar fashion, the only difference being that the scale matrix $\mathbf{B}^{(i)}$ now becomes a weighted sum of squares and cross-products, with current values of q_j ($j = 1, \ldots, J$) serving as the weights (see Equation 29).

References

Aitkin, M., & Longford, N. (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society, Series A, 149,* 1–43.

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

Barnard, G. (1980). Comment on Box (1980). *Journal of the Royal Statistical Society, Series A, 143,* 404–406.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. L. Launer & G. N. Wilkinson (Eds.), *Robusiness in statistics* (pp. 201–236). New York: Academic Press.

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *Journal of the Royal Statistical Society, Series A, 143,* 383–430.

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis.* Reading, MA: Addison-Wesley.

Braun, H., Jones, D., Rubin, D., & Thayer, D. (1983). Empirical Bayes estimation of coefficients in the general linear model from data of deficient rank. *Psychometrika, 48,* 171–181.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin, 101,* 147–158.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage.

Bryk, A. S., Raudenbush, S. W., Seltzer, M., & Congdon, R. (1988). *An introduction to HLM: Computer program and users' guide.* Chicago, IL: University of Chicago, Department of Education.

Carlin, B. (1992). Comment on Morris and Normand (1992). In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. S. Smith (Eds.), *Bayesian statistics 4* (pp. 336–338). New York: Oxford University Press.

Casella, G., & George, E. (1992). Explaining the Gibbs sampler. *American Statistician, 46,* 167–174.

de Leeuw, J., & Kreft, I. (1986). Random coefficient models for multilevel analysis. *Journal of Educational Statistics, 11,* 57–85.

Dellaportas, P., & Smith, A. F. (1993). Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling. *Applied Statistician, 42,* 443–459.

Dempster, A. P. (1983). Comment on Morris (1983). *Journal of the American Statistical Association, 78,* 57–58.

Dempster, A. P. (1987). Comment on Tanner and Wong (1987). *Journal of the American Statistical Association, 82,* 541.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B, 39,* 1–38.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1980). Iteratively reweighted least squares. In P. R. Krishnaiah (Ed.), *Multivariate analysis V* (pp. 35–57). Amsterdam: North-Holland.

Dempster, A. P., Rubin, D. B., & Tsutakawa, R. D. (1981). Estimation in covariance component models. *Journal of the American Statistical Association, 76,* 341–353.

Draper, D. (1995). Inference and hierarchical modeling in the social sciences. *Journal of Educational and Behavioral Statistics, 20,* 115–147.

DuMouchel, W., & Waternaux, C. (1992). Comment on Morris and Normand (1992). In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. S. Smith (Eds.), *Bayesian statistics 4* (pp. 338–341). New York: Oxford University Press.

Efron, B., & Morris, C. (1971). Limiting the risk of Bayes and empirical Bayes estimators—Part I: The Bayes case. *Journal of the American Statistical Association, 66,* 807–814.

Efron, B., & Morris, C. (1972). Limiting the risk of Bayes and empirical Bayes estimators—Part II: The empirical Bayes case. *Journal of the American Statistical Association, 67,* 130–139.

Gelfand, A. E., Hills, S., Racine-Poon, A., & Smith, A. F. M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association, 85,* 972–985.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association, 85,* 398–409.

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

Geyer, C. (1992). Practical Markov chain Monte Carlo. *Statistical Science, 7,* 473–483.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika, 73,* 43–56.

Kackar, R., & Harville, D. (1984). Approximations for standard errors of estimation of fixed and random effects. *Journal of the American Statistical Association, 79,* 853–862.

Kasim, R. (1994). *The application of data augmentation in estimating variance components in multilevel data.* Unpublished doctoral dissertation, Michigan State University, East Lansing.

Kirk, R. (1982). *Experimental design: Procedures for the behavioral sciences.* Monterey, CA: Brooks/Cole.

Laird, N., & Louis, T. (1987). Bootstrapping empirical Bayes estimates to account for sampling variation. *Journal of the American Statistical Association, 82,* 739–757.

Laird, N., & Louis, T. (1989). Empirical Bayes ranking methods. *Journal of Educational Statistics, 14,* 29–46.

Laird, N., & Ware, J. (1982). Random-effects models for longitudinal data. *Biometrics, 38,* 963–974.

Lange, K., Little, R., & Taylor, J. (1989). Robust statistical modeling using the *t* distribution. *Journal of the American Statistical Association, 84,* 881–896.

Lee, P. (1989). *Bayesian statistics: An introduction.* New York: Oxford University Press.

Lee, V., & Bryk, A. (1989). A multilevel model of the social distribution of high school achievement. *Sociology of Education, 62,* 172–192.

Lindley, D. V. (1983). Comment on Morris (1983). *Journal of the American Statistical Association, 78,* 61–62.

Lindley, D. V., & Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B, 34,* 1–41.

Little, R. A. J., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: Wiley and Sons.

Longford, N. (1987). A fast-scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested effects. *Biometrika, 74,* 817–827.

Morris, C. N. (1983). Parametric empirical Bayes inference, theory and applications. *Journal of the American Statistical Association, 78,* 47–65.

Morris, C. N. (1987). Comment on Tanner and Wong (1987). *Journal of the American Statistical Association, 82,* 542–543.

Morris, C. N., & Normand, S. L. (1992). Hierarchical models for combining information and for meta-analyses (with discussion). In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. S. Smith (Eds.), *Bayesian statistics 4* (pp. 321–344). New York: Oxford University Press.

Novick, M., & Jackson, P. (1974). *Statistical methods for educational and psychological research.* New York: McGraw-Hill.

Novick, M., Jackson, P., & Thayer, D. (1971). Bayesian inference and the classical test theory model: Reliability and true scores. *Psychometrika, 36,* 261–288.

Odell, P., & Feiveson, A. (1966). A numerical procedure to generate a sample covariance matrix. *Journal of the American Statistical Association, 61,* 199–203.

Racine-Poon, A. (1992). SAGA: Sample assisted graphical analysis (with discussion). In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. S. Smith (Eds.), *Bayesian statistics 4* (pp. 389–404). New York: Oxford University Press.

Raffe, D. (1991). Assessing the impact of a decentralised initiative: The British Technical and Vocational Education Initiative. In S. W. Raudenbush & J. D. Willms (Eds.), *Schools, classrooms, and pupils: International studies of schooling from a multilevel perspective* (pp. 149–166). San Diego, CA: Academic Press.

Raudenbush, S. W. (1988). Educational applications of hierarchical linear models: A review. *Journal of Educational Statistics, 13,* 85–116.

Raudenbush, S. W. (1992, April). *Hierarchical linear models as generalizations of certain common experimental designs.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Raudenbush, S. W., & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education, 59,* 1–17.

Raudenbush, S. W., Cheong, Y., & Fotiu, R. (1995). Synthesizing cross-national classroom effects data: Alternative models and methods. In M. Binkley, K. Rust, & M. Winglee (Eds.), *Methodological issues in comparative international studies: The case of reading literacy* (pp. 243–286). Washington, DC: National Center for Education Statistics.

Rubin, D. B. (1980). Using empirical Bayes techniques in the Law School Validity Studies. *Journal of the American Statistical Association, 75,* 801–827.

Rubin, D. B. (1981). Estimation in parallel randomized experiments. *Journal of Educational Statistics, 6,* 377–400.

Rubin, D. B. (1983). Iteratively reweighted least squares. In S. Kotz, N. Johnson, & C. Read (Eds.), *Encyclopedia of statistical sciences* (Vol. 4, pp. 272–275). New York: Wiley.

Seltzer, M. (1991). *The use of data augmentation in fitting hierarchical models*

to educational data. Unpublished doctoral dissertation, University of Chicago, Chicago, IL.

Seltzer, M. (1993). Sensitivity analysis for fixed effects in the hierarchical model: A Gibbs sampling approach. *Journal of Educational Statistics, 18,* 207–235.

Smith, W., & Hocking, R. (1972). Algorithm AS53: Wishart variate generator. *Applied Statistician, 21,* 341–345.

Strenio, J., Weisberg, H., & Bryk, A. (1983). Empirical Bayes estimation of individual growth curves and their relationship to covariates. *Biometrics, 39,* 71–86.

Tanner, M. A. (1993). *Tools for statistical inference* (2nd ed.). New York: Springer-Verlag.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association, 82,* 528–550.

Thomas, A., Spiegelhalter, D., & Gilks, W. (1992). A program to perform Bayesian inference using Gibbs sampling. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. S. Smith (Eds.), *Bayesian statistics 4* (pp. 837–842). New York: Oxford University Press.

University of Chicago School Mathematics Project. (1986). *Transition mathematics field study* (Evaluation Report 85/86-TM-2). Chicago, IL: University of Chicago, Department of Education.

Waternaux, C., Laird, N., & Ware, J. (1989). Methods for the analysis of longitudinal data: Blood-lead concentration and cognitive development. *Journal of the American Statistical Association, 84,* 33–41.

West, M. (1984). Outlier models and prior distributions in Bayesian linear regression. *Journal of the Royal Statistical Society, Series B, 46,* 431–439.

Zeger, S., & Karim, M. (1991). Generalized linear models with random effects: A Gibbs sampling approach. *Journal of the American Statistical Association, 86,* 79–86.

Zellner, A. (1971). *An introduction to Bayesian inference in econometrics.* New York: Wiley and Sons.

Authors

MICHAEL H. SELTZER is Assistant Professor, Graduate School of Education and Information Studies, University of California, 405 Hilgard, Los Angeles, CA 90024-1521; mseltzer@ucla.edu. He specializes in the use of hierarchical models in multisite evaluations and studies of change, and in estimation for hierarchical models via the Gibbs sampler.

WING HUNG WONG is Professor, Department of Statistics, The Chinese University of Hong Kong, Shatin, NT, Hong Kong. He specializes in statistical theory and methods, particularly Markov-chain Monte Carlo methods.

ANTHONY S. BRYK is Professor, Department of Education, The University of Chicago, 5835 Kimbark Ave., Chicago, IL 60637. He specializes in the use of hierarchical models in school effectiveness research and studies of growth, and in estimation techniques for hierarchical models.

Received October 21, 1992
Revision received November 21, 1994
Accepted January 6, 1995