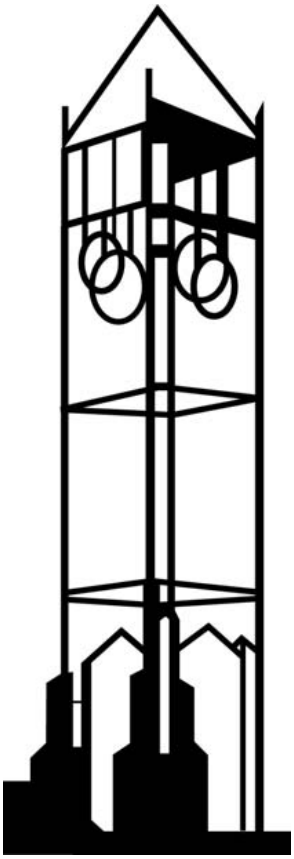
Bayesian Modeling of School Effects Using Hierarchical Models with Smoothing Priors

Mingliang Li, Justin Tobias



Working Paper No. 05004

IOWA STATE UNIVERSITY

Department of Economics
Ames, Iowa, 50011-1070

IOWA STATE UNIVERSITY

Bayesian Modeling of School Effects Using Hierarchical Models with Smoothing Priors

Mingliang Li, Justin Tobias

February 2005

Working Paper # 05004

Department of Economics
Working Papers Series

Ames, Iowa 50011

Bayesian Modeling of School Effects Using Hierarchical Models with Smoothing Priors

Mingliang Li
State University of New York at Buffalo
email: mli3@buffalo.edu

and

Justin L. Tobias
Iowa State University
email: tobiasj@iastate.edu

January 11, 2005

Abstract

We describe a new and flexible framework for modeling school effects. Like previous work in this area, we introduce an empirical model that evaluates school performance on the basis of student level test-score gains. Unlike previous work, however, we introduce a flexible model that relates follow-up student test scores to baseline student test scores and explore for possible nonlinearities in these relationships.

Using data from High School and Beyond (HSB) and adapting the methodology described in Koop and Poirier (2004a), we test and reject the use of specifications that have been frequently used in research and as a basis for policy. We find that nonlinearities are important in the relationship between intake and follow-up achievement, that rankings of schools are sensitive to the model employed, and importantly, that commonly used specifications can give different and potentially misleading assessments of school performance. When estimating our preferred semiparametric specification, we find small but "significant" impacts of some school quality proxies (such as district-level expenditure per pupil) in the production of student achievement.

JEL Codes: C14, I21, J30
Acknowledgements: We would like to thank the two anonymous referees and the editor for helpful comments and suggestions. All errors are our own.

1 Introduction

In recent years we have seen tremendous growth in policies created to hold schools accountable for the production of student achievement.[1] Perhaps the most visible of these policies in the U.S. is the so-called "No Child Left Behind Act" of 2001. Under this act, individual states are given control to come up with their own accountability systems and to define their own criteria for acceptable and meritorious school performance. As an incentive device, the No Child Left Behind Act also *requires* states to provide financial rewards (State Academic Achievement Awards) to those schools exceeding academic achievement targets.

With some exceptions, states (as well as previous studies in this area) evaluate the performances of public schools by looking at growth in test scores, potentially in a variety of subject areas. In what follows, we refer to the first of the two scores used to construct test score growth as *intake achievement* and the subsequent set of test scores as *follow-up achievement*. Generally, a school is viewed as performing well, and potentially may be eligible to receive financial rewards, if follow-up scores significantly exceed intake scores. Perhaps not surprisingly, the specific rules governing accountability systems across U.S. states are quite different, and can be rationalized under different assumptions regarding the relationship between intake and follow-up achievement. These assumptions are extremely important from a policy point of view, as the underlying models provide the foundation for school rankings, rewards allocations and sanctions.

To illustrate varied nature of these programs across states, we can simply look at the accountability programs adopted by (arguably) the three most widely studied states to date: Texas, North Carolina and California. The rewards program implemented in Texas is essentially based on a "value-added" model that uses year-to-year test score *gains* as the evaluation metric. The use of the value added model could be rationalized from a statistical point of view if the relationship between intake and follow-up achievement is linear with a unit slope, whence score gains become the outcome of interest.[2] Unlike the Texas program, North Carolina's ABC's Accountability Model is based on a linear regression model that relates follow-up achievement to initial (intake) achievement.[3] In contrast, California's school accountability plan blends aspects of both the value-added and linear models when determining schools that are deserving of reward funding.[4]

[1] See, e.g., Kane and Staiger (2002) for a recent review.

[2] To be eligible for an award allocation under the Texas Successful Schools Awards System (TSSAS), a school's average test score *gains* in reading and mathematics must rank within the top quartile of a comparison group of 40 similar schools. Specifically, a group of 40 comparison schools is first identified for each school, where the comparison group is determined by finding the 40 schools that are "most similar" to the given school by matching according to various average demographic characteristics. (Details of this procedure can be found, for example, on the website http://www.tea.state.tx.us/perfreport/account/2001/manual.) Student-level test score *gains* on the Texas Learning Index (TLI) are then calculated in reading and mathematics and are aggregated and averaged at the school level. This school-level score in reading and mathematics is then compared to the score received by the 40 comparison schools. If the school ranks within the top quartile of these 40 schools, it becomes eligible for awards under TSSAS.

[3] The formula governing awards allocation in North Carolina (suitably rearranged) is basically a linear regression of follow-up scores in mathematics and reading on the subject intake score, as well as some corrections for "student proficiency" and "regression to the mean." More information about this accountability model is available at http://www.ncpublicschools.org/abcs.

[4] To be eligible for rewards under the Governor's Performance Award Program (GPAP) in California,

Given the growing emphasis on the adoption of school accountability programs and the use of awards systems to reward the most deserving schools, it is clear that proper assessments of school performance are essential. From this simple investigation of existing accountability programs, we also find that there is no universally accepted way to determine those schools that are high-achieving, and observe that most of the programs implemented in practice are based on popular "value-added" and "linear" models.

In this paper we revisit this important issue and present a new and more general way for modeling school effects and assessing school performance. In particular, we recognize that the relationship between intake and follow-up achievement may be nonlinear, as initially low achieving schools and students may be expected to demonstrate large test score gains, while initially high achieving schools and students may only be expected to maintain their level of high achievement.

To flexibly model the relationships between intake and follow-up achievement and explore if such nonlinearities are present, we first recognize that the scores employed in our data set and, in fact, the test scores used in virtually all accountability programs, are discrete in nature. This suggests that one can be fully nonparametric about the relationships between intake and follow-up scores - and thereby nest the widely-used linear and value added models - by simply adding dummy variables for all the possible test score outcomes. We argue, however, that this approach is deficient in the sense that it may potentially overfit the model, that it can (and does in our data) suggest relationships between intake and follow-up scores that are excessively "jumpy", and that it fails to impose intuitive properties like monotonicity in these relationships. To combat these issues we obtain an improved modeling of the conditional mean function by introducing a prior that borrows information from neighboring cells and uses this information to smooth the dummy variable coefficients. We also describe an objective method for determining the amount of smoothing that is most supported by the data, following Koop and Poirier (2004a).

We apply our methods and estimate this model using data from the High School and Beyond (HSB) longitudinal survey. Importantly for our purposes, the sophomore cohort of HSB is given a battery of tests in a variety of different subjects, and then is re-administered these tests two years later during their senior year. This design essentially mimics the evaluation problem currently faced by policy makers and states, wherein students are tested in a variety of subjects at two distinct points in time, and the problem becomes one of determining those schools exhibiting the best performance. In our HSB data, we obtain information on approximately 20,000 students divided among approximately 1,000 different high schools in the United States. We also obtain test score data in 6 different subjects: reading, vocabulary, mathematics, science, writing and civics.

Using the HSB data, we find a number of important results. First, we test and reject

public schools must demonstrate a test score improvement equal to the larger of: (1) 5 Academic Performance Index (API) points (during the 2000-2001 API cycle) and (2) 5 percent of the difference between 800 and the schools base API score. This rule creates a linear spline describing the threshold for awards eligibility where all schools with base API's of at least 700 must demonstrate a 5 point API improvement (i.e., a value-added model, wherein a school receives an award if it *gains* 5 API points), and all schools with base API's less than 700 must increase their score by 5 percent of the difference between 800 and the base API (i.e., a *linear* model given by $y^* = 40 + .95x$, where x is the base score and y^* is the rewards threshold).

the widespread use of linear or value-added models for the HSB data and find important nonlinearities in the relationship between intake and follow-up achievement. Second, we compare *school rankings* under our preferred semiparametric model to those obtained under the value-added and linear models. In general we find that performance assessments obtained in value-added specifications are surprisingly different from those obtained in our preferred semiparametric model, and in particular, the value-added model tends to highly rank schools whose students are initially low-achieving. Third, we consider how the addition of student-level controls such as parental education and income affect the school rankings.[5] We find that the addition of student-level controls does have a reasonable impact on the resulting school rankings and that schools with favorable demographic characteristics are rated most highly in our models. Finally, we also assess the roles of proxies for "school quality" such as teacher education, class size and expenditure per pupil in explaining variation in school performance. Interestingly, we find "significant" effects for some of these variables in our semiparametric model, though the magnitude of the impacts is reasonably small. However, when estimating the "value-added" specification (as often done in the literature), we find no "significant" impacts of the school quality proxies.

The outline of the paper is as follows. In the next section we introduce our semiparametric hierarchical model of achievement growth and discuss the smoothing prior. Section 3 describes the data from the High School and Beyond (HSB) longitudinal study, and empirical results are provided in section 4. The paper concludes with a summary in section 5, and specific details regarding the model and posterior simulator are provided in the appendix.

2 The Model

The model we employ must account for specific features of our data and also allow for the potential to flexibly estimate the relationships between intake and expected follow-up achievement. With respect to features of our data, we need to account for the multilevel clustering structure since we observe multiple test score outcomes for the same student, and students are then clustered into various public high schools.

Before diving into the particular model we use to account for these issues, let us first define some general notation. Let y_{ish} denote the follow-up test score of individual i in subject s in (high) school h,[6] and let x_{ish} similarly denote the baseline score. Since the test score data are discrete, we assume $x_{ish} \in \{x_s^1, x_s^2, \cdots, x_s^{J_s}\}$, with $x_s^1 < x_s^2 < \cdots < x_s^{J_s}$, where J_s denotes the total number of possible test score outcomes on subject test s. We then define

$$D_{ish}^j = \begin{cases} 1 & \text{if } x_{ish} = x_s^j \\ 0 & \text{otherwise} \end{cases}, \quad j = 1, 2, \cdots, J_s \tag{1}$$

[5] This comparison is particularly important as rewards programs in some states (e.g. Texas) explicitly use demographic characteristics in their performance evaluation, while others basically condition on the intake score exclusively (e.g. California).

[6] The High School and Beyond data we use in our empirical analysis contains information on the test score outcomes of high school students, and so we maintain this notation in our modeling.

and let

$$D_{ish} = \left[D_{ish}^1 \; D_{ish}^2 \; \cdots \; D_{ish}^{J_s} \right] \tag{2}$$

be the $1 \times J_s$ *label vector* denoting the test score for individual i in subject s. Finally we let

$$\beta_s = [\beta_s^1 \; \beta_s^2 \; \cdots \; \beta_s^{J_s}]' \tag{3}$$

be the associated $J_s \times 1$ vector of dummy variable coefficients. Importantly, we have sorted the x_s^j so that the dummy variable coefficients are ordered consecutively with the intake test score outcomes.

With this notation in hand, and keeping in mind the hierarchical nature of our problem, we choose to estimate a specification of the following form:[7]

$$y_{ish} = \alpha_{ih} + D_{ish}\beta_s + \epsilon_{ish}, \quad \epsilon_{ish} \overset{ind}{\sim} N(0, \sigma_{ys}^2) \tag{4}$$

$$\alpha_{ih} = \gamma_h + u_{ih}, \quad u_{ih} \overset{iid}{\sim} N(0, \sigma_\alpha^2) \tag{5}$$

$$\gamma_h \overset{iid}{\sim} N(0, \sigma_\gamma^2). \tag{6}$$

This model constitutes a *multilevel hierarchical model*, with α_{ih} representing an individual random effect (given the availability of multiple test score performances for a given individual) and γ_h representing a school random effect (given the clustering of individuals within schools). We are primarily interested in the γ_h parameters, as they can be used to summarize school performance.

The hierarchical approach to this and similar evaluation problems has been undertaken in previous work. For example, Aitkin and Longford (1986) consider the assessment of school effectiveness in educational research studies and Goldstein and Spiegelhalter (1996) describe the statistical issues involved in making quantitative comparisons between institutions in the areas of education and health using hierarchical Bayesian models and Gibbs sampling. Yang *et al.* (2002) introduce the use of multivariate multilevel models and carefully document the need to account for subject selectivity and prior achievement for a large sample of examination results from the U.K. Our goal in this paper is to introduce a new framework for modeling the relationship between base and follow-up achievement, to argue that care needs to be taken to correctly model this relationship, and to further explore how school performance assessments are affected by a variety of other changes in the specification of (4) - (6).

Some features of the model above merit additional discussion. First, our flexible representation in (4) nests the value-added and linear models and allows for different patterns

[7] Of course the follow-up scores (y_{ish}) are also discrete-valued, which is not fully accounted for in our empirical specification. In our application y takes on approximately 100 different values for each subject test so that the continuous approximation is reasonable, and posterior predictive checks suggest that predictions obtained from the continuous model match the observed follow-up data quite well. We exploit the discreteness of the test score variable on the right-hand side as a means to an end: it offers a convenient way to approach nonparametric estimation of the regression function. Estimates for each subject test were also conducted using an ordered probit specification (which does account for the discreteness of y), and were found to be highly similar to those obtained from the model in (4) - (6).

of growth across subject tests.[8] The restricted value-added and linear models have not only provided a basis for policy decisions, as discussed in the introduction, but have also been used extensively in academic research to assess the importance of school characteristics or various school policies.[9] Second, we also allow for the possibility of subject-specific variance parameters, since it is reasonable to expect that the *conditional variability* in test score outcomes may differ by subject test. If the variances differ across subjects, a constant test score improvement should not be rewarded equally across subjects when aggregating results to the school level.

The Smoothing Prior

At this point the model in (4) is just a dummy variable model that is quite flexible and imposes no restrictions on the nature of test score improvements. However, it is reasonable *a priori* to expect some degree of smoothness in the relationships between intake and expected follow-up scores. Direct estimation of (4)-(6) without any additional structure placed on the model may yield estimates that are "undersmoothed" and at odds with our prior beliefs regarding the shape of the regression functions.

To incorporate these features into our model specification, we introduce independent *smoothing priors* on the set of dummy variable coefficients for each subject test. These smoothing priors incorporate our prior belief that adjacent dummy variable coefficients should be similar in value, and the prior will thus tend to "iron out" jumpiness in results obtained from the unrestricted dummy variable model. To be more formal about our prior specification, let $\overline{S}$ denote the total number of subject tests and define the $J_s \times J_s$ first differencing matrix H_s :[10]

$$H_s = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix}, \quad s = 1, 2, \cdots, \overline{S} \tag{7}$$

[8] In the empirical sections, we will show explicitly that our model nests the value-added models and the linear models. In addition, we will also add school quality variables and measurable individual characteristics to our model in the empirical sections, which appears to be a natural extension to our model.

[9] For example, Lee and Smith (1995) and Hanushek, Kain and Rivkin (1998) use value-added specifications while Willms (1985), Hoffer, Greely and Coleman (1985) and Link and Mulligan (1991) use linear specifications. An important exception to this rule is Goldstein and Thomas (1996) who allow for a quartic specification of the intake score.

[10] Note that the intake test scores have been ordered in ascending value, as described before in (1). In the empirical sections, we smooth our regression curves using the second-order differencing. As suggested in Koop and Poirier (2004a), with the second-order differencing and a particular choice of the smoothing prior $V(\eta)$, the smoothing prior approach can match the natural cubic spline approach. In this sense, the second-order differencing appears to have a natural interpretation and it allows the regression curves to be estimated in a very flexible way. More importantly, the smoothness of the regression curves ultimately depends on the smoothing parameter η to be introduced shortly, which governs the smoothness of the differenced parameters. In the empirical sections, we will illustrate in detail how we choose the smoothing prior optimally by finding the η value that maximizes the log marginal likelihood.

and the $J_s \times J_s$ covariance matrix $V_s(\eta_s)$:

$$V_s(\eta_s) = \begin{bmatrix} v_s^2 & 0 \\ 0 & \eta_s I_{J_s-1} \end{bmatrix}, \quad s = 1, 2, \cdots, \overline{S}, \tag{8}$$

where I_M denotes the $M \times M$ identity matrix. We then employ the following priors for the vectors $[\beta_s^1 \quad \beta_s^2 - \beta_s^1 \quad \beta_s^3 - \beta_s^2 \quad \cdots \quad \beta_s^{J_s} - \beta_s^{J_s-1}]'$ $s = 1, 2, \cdots, \overline{S}$, by specifying:

$$H_s \beta_s \overset{ind}{\sim} N[0, \sigma_{ys}^2 V_s(\eta_s)], \quad s = 1, 2, \cdots \overline{S}, \tag{9}$$

or equivalently,

$$\beta_s \overset{ind}{\sim} N\left(0, \sigma_{ys}^2 H_s^{-1} V_s(\eta_s)[H_s^{-1}]'\right) \equiv N(0, \sigma_{ys}^2 \Omega_s) \tag{10}$$

with $\Omega_s \equiv H_s^{-1} V_s(\eta_s)[H_s^{-1}]'$. The parameter β_s^1 acts as an *initial condition*, and our specification in (10) places a prior over that initial condition. The remaining rows of H_s serve to take first-differences of the dummy variable coefficients, and we introduce the potential for smoothing the dummy variable coefficients by centering these first differences over a prior mean of zero. As will be discussed in the following section (following Koop and Poirier (2004a,b)), the values of η_s act as smoothing parameters and will dictate the degrees of smoothness imposed on these regression curves.

Finally, the model specification is completed by adding the following (conjugate) priors:[11]

$$\sigma_{ys}^2 \overset{ind}{\sim} IG(\underline{e}_{1s}, \underline{e}_{2s}), \quad s = 1, 2, \cdots \overline{S} \tag{11}$$

$$\sigma_{\alpha}^2 \sim IG(\underline{a}_1, \underline{a}_2) \tag{12}$$

$$\sigma_{\gamma}^2 \sim IG(\underline{g}_1, \underline{g}_2). \tag{13}$$

The question naturally arises: is this added "smoothing" feature of the model as outlined in (7) - (10) really necessary? We argue strongly in the affirmative. Generally speaking, if the size of the data set at hand is moderate then we would expect imprecise estimation of the dummy variable coefficients, thus yielding erratic estimates of the relationships between base and expected follow up scores. Prior information such as that outlined in (7) - (10) can serve to "smooth out" this excessive jumpiness. Finally, we can simply look into the data to ultimately decide if such smoothing is warranted. Specifically, we can calculate *marginal likelihoods* associated with our smoothed model and a variety of parametric alternatives, including the value-added and linear models and the unrestricted dummy variable specification. These marginal likelihoods will balance the added fit of the unrestricted dummy variable model against the parsimony expressed in the smoothed model. If the data prefer the unrestricted model or simpler parametric alternatives, then seemingly there is little value in the smoothing prior. In our application, however, we find that the smoothed model produces reasonable estimates of the relationships between intake and expected follow-up achievement, and is also strongly favored relative to competing specifications in more formal testing procedures.

[11] In all cases, $IG(a, b)$ denotes the Inverted Gamma density, and is parameterized as in Carlin and Louis (1996, page 326).

2.1 How Does the Smoothing Prior Smooth?

Using the result of Lindley and Smith (1972), one can show that the conditional posterior mean of the coefficient vector β_s is of the following form:[12]

$$E(\beta_s|\eta_s, \alpha, \text{Data}) = \left(D_s' D_s + \Omega_s^{-1}\right)^{-1} D_s' \tilde{y}_s, \tag{14}$$

where

$$D_s = \begin{bmatrix} D_{1sh} \\ D_{2sh} \\ \vdots \\ D_{nsh} \end{bmatrix}, \quad \tilde{y}_s = \begin{bmatrix} y_{1sh} - \alpha_{1h} \\ y_{2sh} - \alpha_{2h} \\ \vdots \\ y_{nsh} - \alpha_{nh} \end{bmatrix} \tag{15}$$

n denotes the number of individuals in the sample and Ω_s is defined as in (10). We note that

$$D_s' D_s = \begin{bmatrix} n_s^1 & 0 & \cdots & 0 \\ 0 & n_s^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & n_s^{J_s} \end{bmatrix} \quad \text{and} \quad D_s' \tilde{y}_s = \begin{bmatrix} n_s^1 \overline{y}_s^1 \\ n_s^2 \overline{y}_s^2 \\ \cdots \\ n_s^{J_s} \overline{y}_s^{J_s} \end{bmatrix}, \tag{16}$$

where n_s^j denotes the number of students with intake test scores equal to x_s^j on subject s and $\overline{y}_s^j$ denotes the average follow-up score (net of individual effects) for those individuals with intake test scores equal to x_s^j, i.e., $\overline{y}_s^j \equiv (n_s^j)^{-1} \sum_{i:x_{ish}=x_s^j} [y_{ish} - \alpha_{ih}]$.

To fix ideas (and without loss of generality), let us make the simplifying assumption that the number of students observed with each initial test score is constant so that $n_s^j = c \; \forall j$. With a bit of work, it follows that the posterior mean in (14) reduces to

$$E(\beta_s|\eta_s, \alpha, \text{Data}) = W_s \overline{y}_s \tag{17}$$

where $\overline{y}_s = [\overline{y}_s^1 \; \overline{y}_s^2 \; \cdots \overline{y}_s^{J_s}]'$, $W_s = W_s(v_s, \eta_s, c)$, and specifically,

$$W_s = c \begin{bmatrix} (v_s^{-2} + \eta_s^{-1} + c) & -\eta_s^{-1} & 0 & 0 & \cdots & 0 & 0 & 0 \\ -\eta_s^{-1} & (2\eta_s^{-1} + c) & -\eta_s^{-1} & 0 & \cdots & 0 & 0 & 0 \\ 0 & -\eta_s^{-1} & (2\eta_s^{-1} + c) & -\eta_s^{-1} & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -\eta_s^{-1} & (2\eta_s^{-1} + c) & -\eta_s^{-1} \\ 0 & 0 & 0 & 0 & \cdots & 0 & -\eta_s^{-1} & \eta_s^{-1} + c \end{bmatrix}^{-1}. \tag{18}$$

In the case of the unrestricted dummy variable model, the weight matrix W_s takes the form $W_s = I_{J_s}$ so that the posterior mean of each element of β_s is simply the average of follow-up outcomes for those students with the given intake score. As $\eta_s \to \infty$, and for v_s^{-2} small, we see that $W_s \to I_{J_s}$, so that the unrestricted dummy variable model results in this limiting case. Conversely, when $\eta_s \to 0$, the coefficients are restricted to be constant, thus "oversmoothing" the model. Without loss of generality, however, (18) shows that expected performance at a particular intake score will be obtained as a weighted average of the average

[12] Note that, conditioned on the individual effects α, the assumptions of our model imply that the coefficients β_s are independent across s. As such, we drop $\{\beta_k\}_{k=1,2,\cdots,\overline{S},k\neq s}$ from the conditioning.

follow-up score at the given intake score and the averages of follow-up scores at neighboring intake scores.

To illustrate the particular way that η_s acts as a smoothing parameter, we calculate the weights as in (18) for a particular case.[13] We set $J_s = 10$ (10 possible test scores), $c = 100$ (100 observations per score) and $v_s^{-2} = 0$ (a diffuse prior over the initial condition). We then determine the weights assigned to the various average follow-up scores for the fifth element of the parameter vector (β_s^5) under a variety of choices for η_s. These are reported in Table 1 below:

Table 1: Weights assigned to Follow-up Averages when Calculating $E(\beta_s^5|\eta_s, \alpha, \text{Data})$ Under Alternate Choices of η_s

Avg. Follow-Up Score	$\eta_s = .0001$	$\eta_s = .001$	$\eta_s = .01$	$\eta_s = .1$	$\eta_s = 1$	$\eta_s = 10$
$\overline{y}_s^1$	.0992	.0900	.0313	.0004	.0000	.0000
$\overline{y}_s^2$	.0997	.0945	.0469	.0023	.0000	.0000
$\overline{y}_s^3$	.1007	.1037	.0860	.0159	.0004	.0000
$\overline{y}_s^4$	.1022	.1181	.1681	.1087	.0185	.0020
$\overline{y}_s^5$	.1042	.1384	.3341	.7454	.9623	.9960
$\overline{y}_s^6$	.1017	.1157	.1673	.1087	.0185	.0020
$\overline{y}_s^7$	.0998	.0987	.0841	.0159	.0004	.0000
$\overline{y}_s^8$	.0983	.0867	.0431	.0023	.0000	.0000
$\overline{y}_s^9$	.0973	.0790	.0235	.0003	.0000	.0000
$\overline{y}_s^{10}$	.0969	.0752	.0157	.0001	.0000	.0000

As shown in Table 1, for small values of η_s (i.e., $\eta_s = .0001$) the weights essentially become uniform and thus we will obtain the same posterior mean for all elements of β_s. For large values of η_s, virtually all the weight is placed on the 5^{th} cell, thus reproducing the dummy variable specification. For intermediate cases, the results are local averages of the neighboring scores, with the weights declining as we move farther away from the 5^{th} cell. In our empirical application, we will calculate marginal likelihoods associated with a variety of η_s and thereby determine and use the amount of smoothing most supported by the data.

2.2 Interpretation of School Effects

The objects of primary interest in this investigation are the school effects γ_h. Though the smoothing feature of our model is important and offers a methodological contribution to this literature, the relationships between intake and follow-up achievement are really only policy-relevant insofar as they are correctly modeled and thus yield accurate assessments of school performance.

Before looking a bit more closely at how our school effects are obtained, let us first introduce some notation. Let θ denote all the parameters in our model and write $\theta =$

[13] Koop and Poirier (2004a) contain a similar discussion, and also discuss the use of a variety of other smoothing priors.

$[\alpha\ \gamma\ \pi]$ to separate the individual random effects α and the school random effects γ from the remaining parameters.

Again, from Lindley and Smith (1972), it immediately follows that

$$E(\gamma_h|\alpha,\pi,\text{Data}) = \frac{n_h\sigma_\gamma^2}{n_h\sigma_\gamma^2+\sigma_\alpha^2}\overline{\alpha}_h, \tag{19}$$

where n_h denotes the number of students in school h and $\overline{\alpha}_h = n_h^{-1}\sum_{i\in h}\alpha_{ih}$ is the sample average of individual effects within school h. This derivation shows that the posterior mean of the effect for school h is simply a weighted average of the average individual effects within the school and the common mean across all schools, which in this case is zero.[14] Thus, schools will be estimated to have large effects if the students within that school are improving more than other students with similar levels of intake achievement.

To be a bit more formal about this claim, we can obtain the conditional mean analogous to (19) but marginalized over the individual effects α.[15] This will enable us to directly see how student-level performances above or below expectation will affect our resulting assessments of school performance. To this end, we let $\alpha_h = [\alpha_{1h}\ \alpha_{2h}\ \cdots\ \alpha_{n_hh}]'$ be the $n_h \times 1$ vector of individual effects from school h. We note

$$\begin{aligned} E(\gamma_h|\pi,\text{Data}) &= E_{\alpha_h|\pi,\text{Data}}\left[E(\gamma_h|\alpha_h,\pi,\text{Data})\right] &(20)\\ &= \left(\frac{\sigma_\gamma^2}{n_h\sigma_\gamma^2+\sigma_\alpha^2}\right)\iota_{n_h}' E(\alpha_h|\pi,\text{Data}), &(21)\end{aligned}$$

where the first line follows by the law of iterated expectations, the second line applies (and slightly rewrites) our formula for the conditional expectation in (19) with ι_m denoting an $m \times 1$ vector of ones.

The expectation $E(\alpha_h|\pi,\text{Data})$ on the right-hand side of (21) does not involve γ_h. Thus, to calculate this expectation we first need to integrate out the school effects from the model in (4) - (6). By substituting γ_h out of (5) and stacking over i within h we obtain the prior $\alpha_h|\sigma_\alpha^2,\sigma_\gamma^2 \sim N(0,\Sigma_h)$, where $\Sigma_h = \sigma_\alpha^2 I_{n_h} + \sigma_\gamma^2\iota_{n_h}\iota_{n_h}'$. Similarly, one can take (4), stack observations first over subject tests s and then over individuals i within school h[16] to obtain the part of the likelihood that depends on α_h.[17] With a bit of work, we obtain the following posterior mean:

$$E(\alpha_h|\pi,\text{Data}) = W_h r_h, \tag{22}$$

[14]Note that for identification purposes the common intercept in (6) is set to zero, as the intercept is implicitly incorporated in the dummy variable specification.

[15]The conditional means $\text{E}(\gamma_h|\pi,\text{Data})$ derived below is marginalized over the individual effects α, but still conditioned on other parameters π. It is worth noting that the conditional means $\text{E}(\gamma_h|\pi,\text{Data})$ are somewhat different from the marginal posterior means $\text{E}(\gamma_h|\text{Data})$ to be estimated from the Gibbs sampler.

[16]Highly similar derivations are reported in the appendix in the description of the posterior simulator, and we do not repeat those derivations here.

[17]Again, note the assumptions of our model imply that observations arising from different schools h are independent from one another. Thus, we do not need to carry along the remaining set of individual effects in the conditioning.

where

$$W_h = \left[\left(\frac{\sum_{s=1}^{\overline{S}} \sigma_{ys}^{-2}}{\overline{S}} \right) I_{n_h} + \frac{\Sigma_h^{-1}}{\overline{S}} \right]^{-1}, \quad r_h = [\overline{y}_{1h}\ \overline{y}_{2h}\ \cdots\ \overline{y}_{n_h h}]' \tag{23}$$

and in the construction of r_h in (23) we have defined

$$\overline{y}_{jh} = \frac{1}{\overline{S}} \sum_{s=1}^{\overline{S}} \frac{y_{jsh} - D_{jsh}\beta_s}{\sigma_{ys}^2} \tag{24}$$

as the average of the "adjusted residuals" on each of the $\overline{S}$ subjects for individual j. Putting (22) together with (21) we obtain the desired mean

$$E(\gamma_h|\pi, \text{Data}) = \left(\frac{\sigma_\gamma^2}{n_h\sigma_\gamma^2 + \sigma_\alpha^2} \right) \iota_{n_h}' W_h r_h. \tag{25}$$

This result directly shows that point estimates of the school effects γ_h are obtained as a weighted average of the average performances of the students within the given school. A school will tend to receive a high ranking if its constituent $\overline{y}_{jh}$ are positive, and from (24), this implies that the students comprising the school have improved more than other students with similar intake ability. These individual performances $\overline{y}_{jh}$ are also a function of $\{\beta_s\}$, and thus *our school performance assessments will clearly be affected by our treatments of the relationships between intake and follow-up scores - the value-added, linear, dummy variable and smoothed dummy variable models can potentially produce dramatically different school rankings.* In our empirical analysis of section 4 we investigate this issue, and find strong evidence that the maintained model can have a significant impact on assessments of school performance. Before discussing these results, however, we first describe the data used in more detail.

3 The Data

We take data from High School and Beyond (HSB) to estimate the models discussed in the previous section. High School and Beyond is an unusually rich longitudinal study, containing detailed information on student achievement, school characteristics and family characteristics of the sampled individuals. Approximately 1,000 randomly selected U.S. high schools participated in the HSB survey, and were chosen to be representative of the population of U.S. high schools. HSB consists of both a sophomore and senior cohort, and approximately 30 students from each cohort were sampled from the participating schools.

In 1980, the base year of the survey, HSB administered cognitive tests in a variety of subjects, including vocabulary, reading, math, science, writing and civics. In 1982, the first-follow up year of the survey, the sophomore cohort was re-administered this test battery, thus allowing us to analyze the production of achievement gains on the various component tests. Since these follow-up test scores are not available for the senior cohort (as they were

high school graduates in 1980), we restrict our attention to the sophomore cohort of HSB.[18]

For each subject test we make use of a formula score provided in the HSB data that corrects scores by imposing penalties for incorrect guessing. For each of the 6 component tests analyzed in this paper (vocabulary, reading, math, science, writing and civics) we obtain a discrete grid of approximately 100 possible test score outcomes. To convert these scores to a common scale, we divide each score by the maximum score, and thus an intake or follow-up score of 1 denotes a perfect score. In general, the transformed scores are then contained in the common interval $[-.2, 1]$ for all subject tests (see, e.g. Figure 1). Though it is possible to receive a negative score, the vast majority of intake and follow-up scores were positive. For example, the percentage of positive intake and follow-up scores in each subject test ranges from 93 to 98 percent.

In some of our models we additionally control for student-level and school-level variables to see how results change with the inclusion of these covariates. The student-level controls we investigate include sex, race, family income, father's education, mother's education and number of siblings. From the school survey in the HSB, we also extract typical proxies for the "quality" of the high school, including average class size, number of books in the school's library, percentage of teachers with at least a Master's or Ph.D. degree and district-level expenditure per pupil. We will examine if these school characteristics play any role in explaining variation in performance across schools in our semiparametric hierarchical model.

The sample sizes we employ in our analysis change as we consider models with additional covariates. In our base model, as described in (4) - (6), which requires only test scores in 1980 and 1982 and individual and school indicator variables, we obtain a final sample of 20,559 students from 953 schools, and thus observe 21.5 students per school on average. When we incorporate student-level controls into the second stage of the hierarchy, the sample size is reduced to 13,404 students from 941 schools due to missing data on family characteristics. Finally, in our models which include both individual-level and school-level covariates, we obtain a final sample of 8,983 students from 599 schools.

4 Empirical Results

In this section we look into the HSB data to address the following key questions: (1) Do the data support the use of the semiparametric smoothing methods over popular alternatives like value-added and linear models? Are nonlinearities in the relationship between intake and follow-up achievement important? (2) Are key model outputs such as assessments of school performance (i.e., school rankings) sensitive to the model employed? (3) Does accounting for student-level controls like parental education and family income affect our school rankings? (4) Do proxies for "school quality" like class sizes, teacher education and

[18] We do not exclude the dropouts from the sophomore cohort. Importantly, dropouts only comprise a small portion of the entire sample. Moreover, most dropouts spend considerably long periods of time in their high schools before they drop out. Therefore, it is still reasonable to assume that dropouts and in-school students from the same school share the same school-specific effect. In addition, the results of our paper are robust to the exclusion of the dropouts from the sample.

expenditure per pupil explain variation in school performance in our semiparametric model? (5) Do we obtain the same impact of the "school quality" variables when estimating the widely-used value-added and linear models?

The models we employ are fit using the *Gibbs sampler*. In the appendix of this paper we describe our algorithm for fitting the model in (4) - (6) by employing a *blocking step* wherein the regression parameters $\{\beta_s\}$, individual effects $\{\alpha_{ih}\}$ and school effects $\{\gamma_h\}$ are drawn in a single block. This helps to mitigate autocorrelation in our parameter chains. Our posterior means and other parameters of interest are calculated after running the sampler for 10,000 iterations and discarding the first 1,000 of these simulations as the burn-in period. Some of our models add covariates to the second and third stages of the hierarchy, and these models can be fit using simple modifications of the algorithm described in detail in the appendix.

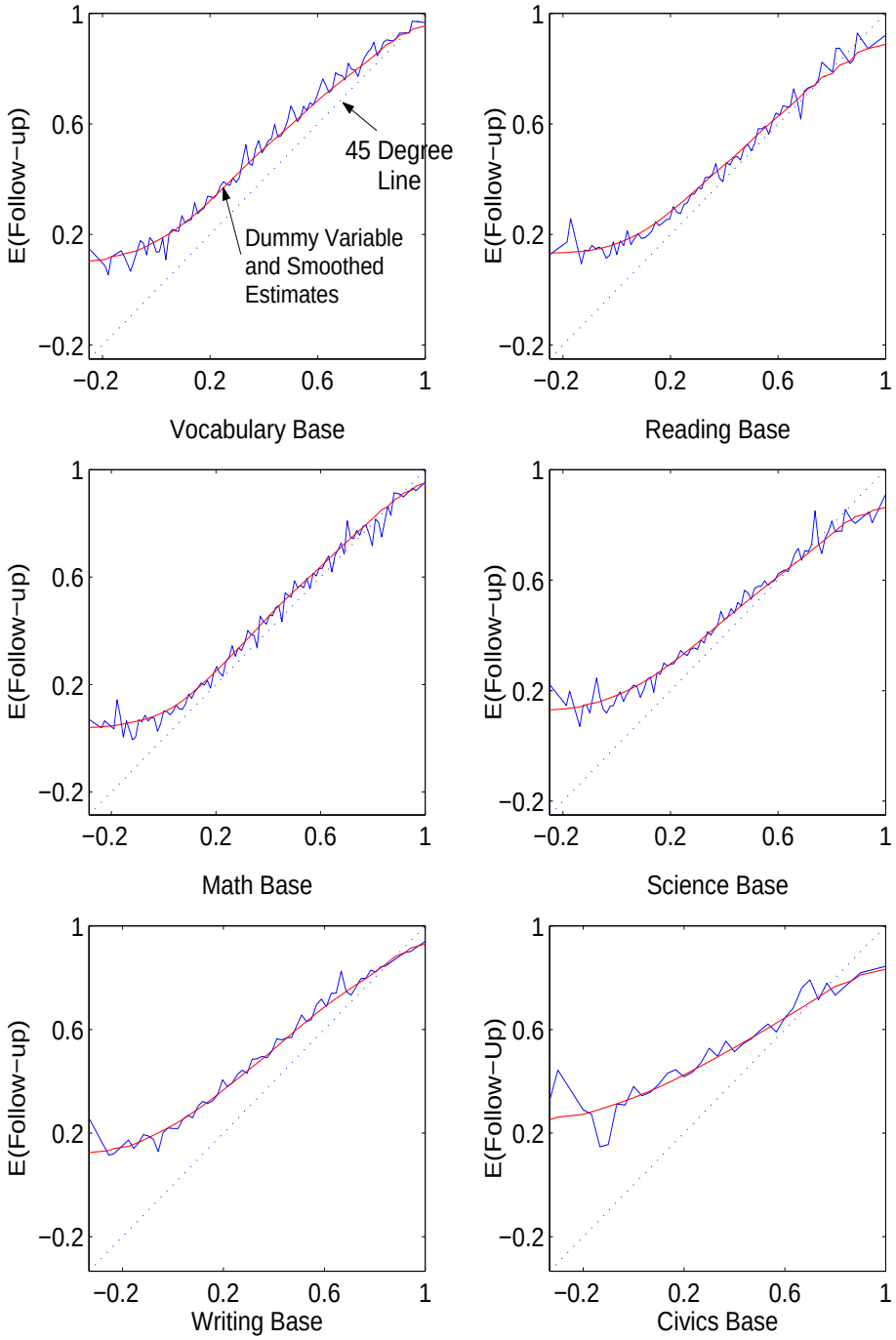
We specify proper yet reasonably "flat" priors for all the parameters in our model. Specifically, for the variance parameters $\{\sigma^2_{ys}\}$, σ^2_α and σ^2_γ, we make the choices: $\underline{e}_{1s} = 1$, $\underline{e}_{2s} = 1$, $\underline{a}_1 = 1$, $\underline{a}_2 = 1$, $\underline{g}_1 = 1$ and $\underline{g}_2 = 1$, and found that setting all these parameters to 10 or .1 produced virtually identical results. For the variance parameter governing the initial condition in (8), we specify $v^2_s = 10,000 \quad \forall s$. Finally, for our most general model specifications we also include student and school level covariates. When these are included, we center the coefficients around a prior mean of zero with a prior covariance matrix equal to $1,000 I_k$. These specifications are chosen to be vague so that the data information is predominant, and results were not found to be sensitive to moderate changes in these priors.

4.1 A Series of Test-Specific Models

To fix ideas and illustrate the potential advantages of "smoothing" we begin by treating each of the six subjects (vocabulary, reading, math, science, writing and civics) independently. Specifically, we estimate six different linear regression models where the follow up scores of each test are regressed on an exhaustive set of dummy variables. We then place a smoothing prior on these dummy variable coefficients as in (10). These simplified models are not hierarchical, and do not account for the group structure of our HSB data. However, estimation of these simplified models provides a clear way to visualize inadequacies with the unrestricted dummy variable approach, and also reveals how our preferred method overcomes these deficiencies.

We present in Figure 1 the results of this estimation exercise. In Figure 1 we plot the expected follow-up scores for each test against the value of intake achievement using the unrestricted dummy variable regression model. For each of the six component tests we also plot a dotted 45 degree line, which represents the case where intake and expected follow-up scores coincide. In general, expected follow-up scores lie above the 45 degree line in all of our component tests, and only fall below this line at the far right-tail of the intake achievement distribution. This indicates that the vast majority of students have improved, and a closer inspection of Figure 1 shows that those individuals with low intake test scores tend to demonstrate the most improvement. *The fact that the "production" of test score improvements is not constant over the intake support is evidence against the popular*

Figure 1: Expected Follow-up Test Score Given Base Year Test Score: Dummy Variable and Smoothed models



value-added model, where it is implicitly assumed that achievement gains are treated equally regardless of the intake test score.

For our purposes, what may be most important to note is that estimates from the dummy variable model in Figure 1 are quite erratic, while we expect them to be smooth and non-decreasing. Even abstracting from the tails, where the "bumpiness" of the curves seems most pronounced and the data are most scarce, the regression curves still appear to be fluctuating considerably over the interior of the support. In general, our prior beliefs are that these curves should be smooth, and that those with higher intake scores should also have higher expected follow-up scores. Thus, the introduction of the smoothing prior in (10) seems potentially advantageous, as it will help to impose these conditions on our regression functions.

In Figure 1 we also present our semiparametric results that smooth the dummy variable coefficients. For each of the 6 component tests, we choose the smoothing prior optimally by finding the η value that maximizes the log marginal likelihood.[19] As you can see, the smoothed regression functions mimic the overall trends suggested by the dummy variable model quite well, and even capture the nonlinearities suggested by the dummy variable model in the tails of the regression functions. These smoothed regression curves also embody our prior beliefs that the regression function should be increasing, though it is important to note that we have not formally imposed this condition through our prior. Finally, we also note that the nature of test score improvements seems different across tests (particularly for civics), and thus one should not simply aggregate these test scores when estimating school performance.

To be a bit more formal about our model comparison methods, we present in Table 2 a table of *log marginal likelihoods* for a variety of models in each subject area. Marginal likelihoods are widely used in Bayesian testing and their use arises from the observation that for any two competing models M_1 and M_2:

$$\frac{p(M_1|y)}{p(M_2|y)} = \left(\frac{p(y|M_1)}{p(y|M_2)}\right)\frac{p(M_1)}{p(M_2)}. \tag{26}$$

The left-hand side of (26) gives the *posterior odds* of Model 1 in favor of Model 2, and the ratio $p(M_1)/p(M_2)$ is the *prior odds ratio*, typically taken to be unity. The expression in parentheses following the equality in (26) is the *Bayes factor* or the ratio of marginal likelihoods, with $p(y|M_i)$ denoting the marginal likelihood for Model i. Thus, under equal prior odds, posterior odds ratios can be obtained by exponentiating the difference between the log marginal likelihoods.

In Table 2 we report the log marginal likelihoods for a variety of models. These include the *value-added model* where the dependent variable is the change in test score, the *linear model* which regresses the follow-up score on the intake score, the unrestricted *dummy variable model* and the *smoothed* dummy variable model for an optimally chosen smoothing

[19] These cross-sectional models fit into the framework of a standard linear regression model with a natural conjugate prior for β (see (10)) and a conjugate inverse gamma prior for σ_y^2. Thus, marginal likelihoods can be obtained *analytically* for a given value of η, and so one can select the value of η that maximizes the marginal likelihood via a grid search.

Table 2: Test-Specific Log Marginal Likelihoods for a Variety of Models

model/ subject	value added	dummy variable	linear	smoothed
vocabulary	-39,842	-39,375	-39,220	-39,162
reading	-42,272	-41,559	-41,429	-41,360
math	-40,620	-40,489	-40,345	-40,244
science	-39,716	-38,626	-38,458	-38,417
writing	-45,184	-43,796	-43,664	-43,634
civics	-49,691	-47,256	-47,179	-47,174

parameter. Interestingly, for all of the various subject tests, the rank-ordering of the models is the same - the semiparametric model is most preferred by the data, followed by the linear model, and the value-added specification is least preferred by the data.

4.2 The Full Hierarchical Model

The previous section focused on each subject test independently to illustrate the potential benefits of the smoothing prior. However, these analyses did not account for the panel structure of our data, and thus the models employed were not rich enough to make assessments regarding school performance. In this section, we take up estimation of the general hierarchical specification given in (4) - (6).

We choose to simplify our model by making the restriction $\eta_s = \eta$, thus limiting the model to one smoothing parameter that will be used in the priors for the dummy variable coefficient vectors for all the component tests. For this hierarchical model, marginal likelihoods are not available in closed form, making it difficult to implement an empirical Bayes procedure to search over six dimensions for an optimal vector of smoothing parameters.[20] In addition, we note that each component test is measured on essentially the same scale, so there is no compelling reason to require the use of test-specific smoothing parameters.

Again, we select the optimal value of the smoothing parameter by calculating the marginal likelihoods for various models indexed by different values of η. We present these log marginal likelihood calculations in Table 3.

Table 3 reveals the U-shaped nature of these marginal likelihoods when plotted over the support of η. Using these results, we find $\eta = 2 \times 10^{-5}$ to be the one yielding the highest value of the log marginal likelihood, and thus we use this value in our remaining

[20] To investigate how this assumption impacted our results, we implemented a different algorithm, where $\{\eta_s\}$ were included as parameters in our sampling routine. Results from that analysis were very similar to those presented here. We calculate the marginal likelihoods using the Laplace-Metropolis method discussed in Raftery (1996). When performing the test-by-test analysis of the previous section, marginal likelihoods could be obtained analytically, and we obtained exactly the same results when using the Laplace-Metropolis method. When applying the estimator, we first integrated out the individual and school random effects and the set of dummy variable coefficients for each subject test.

Table 3: Log Marginal Likelihoods for Hierarchical Model With Different Values of Smoothing parameter η

Parameter Value	Log Marginal Likelihood
$\eta = 2 \times 10^{-1}$	-240,377
$\eta = 2 \times 10^{-2}$	-240,169
$\eta = 2 \times 10^{-3}$	-240,052
$\eta = 2 \times 10^{-4}$	-239,970
$\eta = 2 \times 10^{-5}$	-239,935
$\eta = 2 \times 10^{-6}$	-239,969
$\eta = 2 \times 10^{-7}$	-240,107
$\eta = 2 \times 10^{-8}$	-240,285
$\eta = 2 \times 10^{-9}$	-240,336

calculations.[21]

With the optimal value of the smoothing parameter in hand, we now wish to compare our *smoothed semiparametric*[22] model to its competitors. We consider 6 competing models in total, with 5 of these being restricted versions of a linear specification of (4), written as follows:

$$y_{ish} = \alpha_{ih} + [\beta_s^0 + \beta_s^1 x_{ish}] + \epsilon_{ish}, \quad \epsilon_{ish} \overset{ind}{\sim} N(0, \sigma_{ys}^2). \tag{27}$$

Two of the 5 competing models based on (27) are *value-added* models which restrict the slope coefficient in the linear model to unity: $\beta_s^1 = 1$, so that the dependent variable is the *gain* score $y_{ish} - x_{ish}$. In one of our value added models (**VA1**) we also impose equality of intercepts across subject tests: $\beta_s^0 = \beta^0$.

The remaining 3 competitors based on (27) are linear models. The most restrictive of these, denoted **L1**, restricts intercepts and slopes to be the same across tests ($\beta_s^0 = \beta^0$ and $\beta_s^1 = \beta^1$). The linear model **L2** restricts only the slopes to be constant: $\beta_s^1 = \beta^1$, and the least restrictive model, denoted **L3**, allows intercepts and slopes to vary across tests and is given by (27).

Our final competitor is the unrestricted dummy variable model which generalizes the linear specification in (27). This specification, denoted **DUM**, is given in (4) and results in the limiting case as the smoothing parameter $\eta \to \infty$. Of course, we also calculate the log marginal likelihood associated with the smoothed semiparametric model, denoted **SEM**, using $\eta = 2 \times 10^{-5}$. The results of these calculations are given in Table 4.

As seen from the table, *value-added models receive the least support from the data, and our semiparametric model is most supported by the data.* Among those models assuming linearity, it is clear that one should not treat each subject test identically, as the most supported linear model is the one allowing for test-specific intercepts and slopes. This

[21] In practice, the use of neighboring values of η produced virtually identical results.

[22] We call this model a semiparametric model since it involves a nonparametric treatment of the relationships between intake and follow-up scores and maintains parametric assumptions about other aspects of the model.

Table 4: Log Marginal Likelihoods for Alternate Models: Full Hierarchical Model

Model	Restrictions on (27) or (4)	Log Marginal Likelihood
VA1	Eqn. (27) with $\beta_s^0 = \beta^0$, $\beta_s^1 = 1$	-255,815
VA2	Eqn. (27) with $\beta_s^1 = 1$	-254,644
L1	Eqn. (27) with $\beta_s^0 = \beta^0$, $\beta_s^1 = \beta^1$	-245,240
L2	Eqn. (27) with $\beta_s^1 = \beta^1$	-241,612
DUM	Eqn. (4) with $\eta = \infty$	-241,080
L3	Eqn. (27)	-240,234
SEM	Eqn. (4) with $\eta = 2 \times 10^{-5}$	-239,938

suggests the nature of achievement growth varies by subject, and as such, when attempting to assess school or student performance, one needs to estimate a model that does not involve simple aggregation of scores across subjects. Finally, we again note that the data prefer flexibility in the estimation of the test score relationships up to the flexibility offered by the smoothed semiparametric model; the more flexible and parameter-rich dummy variable model did not receive as much support from the data as the smoothed version.

4.3 Do Student-Level Controls Change Results?

In this section we investigate the sensitivity of results to the addition of student level covariates at the second-stage of the hierarchy (equation (5)). This exercise is important, as some states explicitly make use of demographic characteristics like racial and ethnic composition when defining a group of comparison schools for awards eligibility (e.g., Texas), while other states essentially condition on intake achievement exclusively (e.g., California).[23] In addition to these basic demographic variables, characteristics like family income and parental education can also play a potentially important role in the production of student achievement *growth*. If this is true, then schools whose students come primarily from wealthy and well-educated families will tend to be ranked highly in our hierarchical model. This high ranking may not be due to a true school effect, but instead, may arise from the contribution of family background characteristics that have gone unmodeled.

To this end we now include a male indicator, a white indicator, family income, highest grade completed by the respondent's mother and father and number of siblings as covariates in (5). We stack these individual-level observables into a vector z_{ih} and write (5) as:

$$\alpha_{ih} = \gamma_h + z_{ih}\delta + u_{ih}, \quad u_{ih} \overset{iid}{\sim} N(0, \sigma_\alpha^2). \tag{28}$$

The parameters δ can be estimated in a straight-forward generalization of the algorithm provided in the appendix.[24] With this specification, the conditional posterior mean in (19)

[23] To be awards eligible in California, all numerically significant subgroups within the school must also demonstrate adequate API improvement. California does not, however, base its awards allocation decision by first "matching" the given school with schools with similar demographic characteristics. For all groups, the decision is made exclusively on changes in API scores.

[24] For the sake of brevity, we do not provide a table of estimation results here, but will provide such a table in the following section.

now becomes

$$E(\gamma_h|\alpha,\theta,\text{Data}) = \frac{n_h\sigma_\gamma^2}{n_h\sigma_\gamma^2+\sigma_\alpha^2}\overline{\tilde{\alpha}}_h, \tag{29}$$

where $\overline{\tilde{\alpha}}_h = n_h^{-1}\sum_{i\in h}(\alpha_{ih}-z_{ih}\delta)$. Thus, if the characteristics z_{ih} contribute positively to the individual effects α_{ih} (e.g., coming from a family with high parental education and income also improves your test score *growth*), these effects will tend to be "subtracted off" when estimating school-level effects based on the generalized model in (28).

To investigate if our assessments of school performance are sensitive to the inclusion of individual-level characteristics, we entertain the 7 models described in Table 4, and for each of these we obtain *school rankings* with and without the inclusion of the individual-level characteristics z_{ih}. Specifically, for each of the 7 models enumerated in Table 4, we obtain a vector of school random effects $\gamma = [\gamma_1\ \gamma_2\ \cdots\ \gamma_H]$ for each post-convergence draw from the Gibbs sampler. For each iteration, the elements of this vector can be ordered to form a ranking of the various schools. To see how similar these performance rankings are across models that include or omit the characteristics z_{ih}, we appeal to the *rank correlation*. When there are no ties (as is the case here) the rank correlation r can be calculated as

$$r = 1 - \frac{6\sum_{i=1}^{\overline{S}}(M_i^1 - M_i^2)^2}{\overline{S}(\overline{S}^2-1)}, \tag{30}$$

where $\overline{S}$ denotes the total number of schools, M_i^1 denotes the ranking of school i under a particular model in Table 4 without including z_{ih}, and M_i^2 denotes the ranking of school i under that same model specification upon including z_{ih} as in (28).

For a particular model chosen from Table 4, and for every post-convergence draw from our sampler, we calculate the (rank) correlation between the vector of school rankings obtained when including or omitting z_{ih} in (28). Point estimates of the similarity in rankings are then obtained by taking averages of the resulting series of rank correlations. Since this procedure essentially simulates the posterior distribution of the rank correlation, we can also calculate its posterior standard deviation. If the school effects are precisely estimated, then the vector of rankings will not change significantly from iteration to iteration, and thus the calculated rank correlations will remain essentially constant across iterations. The posterior standard deviation thus allows us to quantify our degree of uncertainty surrounding the value of the rank correlation which arises from uncertainty in the school rankings themselves.

As shown in the Table 5, the rank-ordering of schools is reasonably affected by the inclusion of student-level characteristics z_{ih}, and the magnitudes of the impacts are virtually identical across model specifications. Specifically, our point estimates of the rank correlations are all close to .76 with small posterior standard deviations, indicating that including or omitting the individual-level characteristics z_{ih} produces a similar overall effect on school rankings across our model specifications.

To investigate the relationship between family background characteristics and our school rankings in more detail, we present additional information in Table 6. We begin

Table 5: Rank Correlations Between School Rankings With and Without the Inclusion of Student-Level Covariates

Model	Restrictions on (4) or (27)	post. mean	post. std.	post. prob. positive
VA1	Eqn. (27) with $\beta_s^0 = \beta^0$, $\beta_s^1 = 1$	0.7502	0.0167	1
VA2	Eqn. (27) with $\beta_s^1 = 1$	0.7542	0.01632	1
L1	Eqn. (27) with $\beta_s^0 = \beta^0$, $\beta_s^1 = \beta^1$	0.7568	0.01584	1
L2	Eqn. (27) with $\beta_s^1 = \beta^1$	0.769	0.0153	1
L3	Eqn. (27)	0.7686	0.01504	1
DUM	Eqn. (4) with $\eta = \infty$	0.7642	0.01471	1
SEM	Eqn. (4) with $\eta = 2 \times 10^{-5}$	0.765	0.01459	1

by analyzing our original semiparametric model, as outlined in (4) - (6), without the inclusion of the covariates z_{ih}. We then break the vector of school rankings obtained from this model into deciles, and calculate the average values of family characteristics for the schools within each decile. Our intuition is that highly ranked schools will also be the ones with "favorable" demographic characteristics, as these characteristics are likely to play a significant role in achievement growth. The results of this analysis are reported in Table 6 for parental income, parental education and number of siblings.

Table 6: Average Family Characteristics by Decile of School Rankings. Semiparametric Specification in (4) - (6)

	Family Income/ $ 1,000		Father Education		Mother Education		Siblings	
Decile	Mean	Std	Mean	Std	Mean	Std	Mean	Std
1	26.4	(.390)	15.0	(.113)	14.1	(.087)	2.47	(.044)
2	23.7	(.519)	14.2	(.150)	13.4	(.111)	2.60	(.061)
3	23.0	(.537)	13.8	(.148)	13.2	(.109)	2.68	(.064)
4	22.3	(.532)	13.5	(.141)	13.0	(.107)	2.73	(.066)
5	21.4	(.506)	13.2	(.132)	12.8	(.104)	2.79	(.067)
6	20.8	(.484)	13.1	(.120)	12.7	(.098)	2.85	(.070)
7	20.5	(.475)	12.9	(.119)	12.6	(.098)	2.89	(.070)
8	20.2	(.472)	12.8	(.111)	12.5	(.097)	2.91	(.070)
9	19.2	(.483)	12.4	(.099)	12.2	(.091)	3.03	(.074)
10	16.7	(.363)	12.0	(.062)	12.0	(.061)	3.36	(.059)

As shown in Table 6, family characteristics are *strongly* related to the rankings of schools from our semiparametric model. In fact, the average family characteristics are *monotonic* with the decile rankings. Schools in the highest deciles are comprised of students coming from the wealthiest and most educated families with fewest siblings. This suggests that rankings which do not account for these family characteristics may confound actual school performance with demographic characteristics of the students attending those schools.

4.4 Does School Quality Matter?

In the previous subsection we added student-level controls to the second-stage of the hierarchy as in (28). We now investigate the roles of proxies for school "quality" (including class size, number of books in the school's library, percentage of teachers with at least a Master's Degree and district-level expenditure per pupil) in explaining variation in performances across schools. We stack these school-level observables into a vector q_h and write (6) as:[25]

$$\gamma_h = q_h\pi + v_h, \quad v_h \overset{iid}{\sim} N(0, \sigma_\gamma^2). \tag{31}$$

The model examined in this section is then given by (4), (28) and (31) and contains student (z_{ih}) and school (q_h) level observables. It is important to recognize that the results we obtain here are based on our preferred semiparametric specification, which has not been used in previous work, and can potentially give different results than those based on linear and value-added specifications. We present coefficient posterior means, standard deviations and probabilities of being positive from this semiparametric model in Table 7.

Table 7: Posterior means, Standard deviations and Probabilities of being Positive: Smoothed Semiparametric Model with Individual and School Covariates

	Parameter/ Variable	Post. Mean	Post. Std.	Post. Prob. Positive
Individual Covariates	Male	0.00196	0.00249	0.789
	White	0.0671	0.00339	1
	Family income/ $1000	0.000698	0.000124	1
	Father education	0.00503	0.000376	1
	Mother education	0.0033	0.00043	1
	Number of siblings	-0.00458	0.000726	0
School Covariates	Class size	-3.83e-005	0.000191	0.411
	Books in library/ 1,000	0.000678	0.00045	0.94
	Perc. Teacher MA/Ph.D.	0.0164	0.0182	0.816
	District expenditure per pupil / $1,000	0.0125	0.00606	0.986
Variance Parameters	σ_α^2	0.0102	0.000228	1
	σ_γ^2	0.00855	0.00135	1
	$\sigma_{Vocabulary}^2$	0.0161	0.000291	1
	$\sigma_{Reading}^2$	0.0204	0.000353	1
	σ_{Math}^2	0.0195	0.000333	1
	$\sigma_{Science}^2$	0.0154	0.000281	1
	$\sigma_{Writing}^2$	0.0265	0.000449	1
	σ_{Civics}^2	0.0388	0.000619	1

Table 7 confirms the results of the previous section and shows that family income, parental education and family size are strongly related to follow-up achievement even after flexibly controlling for intake achievement. Interestingly, we also find reasonably "significant" effects of proxies for school quality, as district expenditure per pupil and number of

[25] We use π to denote the coefficients of q_h.

books in the school's library[26] have high posterior probabilities of being positive.[27] Finally, the coefficients associated with all of our proxies for school quality have the expected signs.

To get a different sense of how school characteristics impact student performance, we conduct the following simulation exercise. Consider the predictive test score outcomes for an out-of sample individual, and denote these test scores as y_{fsh}, with f denoting a future, as yet unobserved quantity. Assuming the model in (4), (28) and (31) applies to this out of sample individual, we obtain[28]

$$y_{fsh}|\theta_{-\alpha,\gamma}, q_h, z_{ih}, \{D_{fsh}\}, \text{Data} \sim N\left(q_h\pi + z_{fh}\delta + D_{fsh}\beta_s, \sigma_{ys}^2 + \sigma_\alpha^2 + \sigma_\gamma^2\right), \tag{32}$$

where we have integrated out the individual and school random effects. We can then calculate the predictive density associated with the average follow-up score for this individual, denoted $\overline{y}_{fh}$, where

$$\overline{y}_{fh} = \frac{1}{\overline{S}}\sum_{s=1}^{\overline{S}} y_{fsh}.$$

Using (32) we obtain

$$\overline{y}_{fh}|\theta_{-\alpha,\gamma}, q_h, z_{ih}, \{D_{fsh}\}, \text{Data} \sim N\left(q_h\pi + z_{fh}\delta + \overline{D_{fsh}\beta_s}, \overline{\sigma_{ys}^2}/\overline{S} + \sigma_\gamma^2 + \sigma_\alpha^2\right), \tag{33}$$

where $\overline{D_{fsh}\beta_s} = (1/\overline{S})\sum_s D_{fsh}\beta_s$ and $\overline{\sigma_{ys}^2} = (1/\overline{S})\sum_s \sigma_{ys}^2$. We calculate the mean of $\overline{y}_{fh}$ upon fixing each D_{fsh} to the median intake score on subject s and the elements of z to their mean values.[29] We then estimate the mean of (33) for a variety of values of the school quality variables. Specifically, we set each of them equal to their mean values and also set each of them (collectively[30]) to one and two standard deviations above and below their mean values. We find that the posterior means of (33) were .51, .53, .55, .56 and .58 when the quality variables were set equal to $E(q) + c\text{Std(q)}$ for $c = -2,\ -1,\ 0,\ 1$ and 2, respectively. A student-level average follow-up score of .58 would place the student at the 52^{nd} percentile of the observed follow-up score achievement distribution, while a score of .51 would place that student at approximately the 41^{st} percentile. *So, seemingly, large changes in school quality characteristics do have a moderate impact on expected follow-up performance.*

We can look at this problem a bit differently and calculate the posterior predictive probability that a student coming from a school with characteristics 2 standard deviations above the mean will receive a higher average follow-up score than a student with characteristics

[26]This variable also proxies for school size effects, which has been demonstrated in previous work to have some effect on student outcomes, e.g., Betts (1995).

[27]Given our specification of the dependent variable, we can interpret these coefficients as percentage changes in the maximum formula score corresponding to a unit change in the covariate. For example, increasing expenditure per pupil by $1,000 increases expected follow-up scores by about 1.25 percent of the maximum score. Similarly, a 10 point increase in the percentage of teachers in the school with at least an MA increases expected follow-up scores by about .164 percent of the maximum score. These calculations show that the magnitudes of the school quality effects are rather small, though in some cases still "significant."

[28]As in the appendix, we use θ_x to denote all parameters in the model other than x.

[29]We set the white and male dummy variables equal to one.

[30]When we collectively set each of the school quality variables to one and two standard deviations above and below their mean values, we change them in the same direction except for the class size variable simply because the class size variable has an opposite effect on the achievement growth compared with other school quality variables.

2 standard deviations below the mean. Fixing the covariates other than q_h at mean values and using our model in (33), this probability reduces to $\Phi\left([(q_h^2 - q_h^{-2})\pi]/\sqrt{2\sigma^2}\right)$, where q_h^2 denotes the quality values 2 standard deviations above the mean, q_h^{-2} denotes those two standard deviations below the mean, and $\sigma^2 \equiv \overline{\sigma_{ys}^2}/\overline{S} + \sigma_\gamma^2 + \sigma_\alpha^2$. At posterior mean values, we calculate this probability to be approximately .63. The overall .07 point increase in expected follow-up scores as a result of the school quality improvements is rather modest relative to the standard deviation of $\overline{y}_{fh}$ which is approximately .15. *Thus, even though we find "significant" effects of some of the school quality proxies, our analysis suggests there is a reasonably large chance that a student from a quality-poor school will still perform better than a student from a school comparably rich in quality.*

School Quality and Value-Added Models

Finally, it is useful to repeat the analysis of Table 7 using the popular value-added specification. In this way, we investigate if our treatment of the relationship between intake and follow-up achievement has any impact on our assessments of the importance of proxies for school quality. Results obtained using the value-added model **VA2** are reported in Table 8.

Table 8: Posterior means, Standard deviations and Probabilities of being Positive: Value-Added Model with Individual and School Covariates

	Parameter/ Variable	Post. Mean	Post. Std.	Post. Prob. Positive
Individual Covariates	Male	0.00191	0.0022	0.815
	White	0.0088	0.00307	0.998
	Family income/ $1000	9.16e-005	0.000113	0.798
	Father education	0.000874	0.000326	0.997
	Mother education	0.000397	0.000372	0.853
	Number of siblings	-0.000266	0.000667	0.344
School Covariates	Class size	-6.46e-005	0.000161	0.343
	Books in library/ 1,000	6.69e-005	0.000373	0.56
	Perc. teacher MA/Ph.D	-0.0087	0.0147	0.279
	District expenditure per pupil/ $1,000	0.00115	0.00449	0.607
Variance Parameters	σ_α^2	0.00567	0.000145	1
	σ_γ^2	0.0062	0.000377	1
	$\sigma_{Vocabulary}^2$	0.0233	0.000392	1
	$\sigma_{Reading}^2$	0.0297	0.000485	1
	σ_{Math}^2	0.0245	0.000406	1
	$\sigma_{Science}^2$	0.0231	0.000393	1
	$\sigma_{Writing}^2$	0.0383	0.000617	1
	σ_{Civics}^2	0.0666	0.00105	1

The most important piece of information to take away from Table 8 is that we do not see "significant" effects for any of the school quality variables in the value-added specification. In addition, the coefficient on the teacher education variable is negative, contrary to our expectation, and the magnitudes of the impacts are also significantly reduced relative to those in Table 7. To summarize, *use of the value-added specification led us to reach*

very different conclusions regarding the "significance" of the school quality variables: in our preferred semiparametric model, we find some evidence of school quality effects while the popular value-added specification does not reveal such effects. We rationalize this result by observing (see Figure 1) that the value-added specification does not correctly model the relationship between intake and follow-up achievement. More than this, the value-added model will tend to highly rank schools that are initially low achieving and thereby evaluate the importance of school quality variables off of these arguably misleading performance assessments. We discuss this point in greater detail in the following section.

4.5 Are Assessments of School Performance Sensitive to the Conditional Mean Specification?

In the previous sections we argued that our semiparametric specification was preferred over various alternatives, and also noted that including student-level demographic information had an impact on our assessments of school performance. In this section we investigate how the specification of the relationships between intake and follow-up achievement affect our results. Despite the statistical preference of the smoothed model, policy-relevant quantities such as school rankings may not be sensitive to the use of the value-added, linear, dummy variable, or smoothed models.

As in section 4.3, we appeal to the *rank correlation* to quantify the degree of similarity between school rankings obtained from various specifications. We consider the 7 models listed in Table 4 and calculate the pairwise correlations between school rankings from each model. We report these calculations in Table 9.

Table 9: Similarity of School Rankings: Rank Correlations between Ordered School Effects Across Competing Specifications

Model	Restrictions on (27) or (4)	**VA1**	**VA2**	**L1**	**L2**	**L3**	**DUM**	**SEM**
VA1	Eqn. (27) with $\beta_s^0 = \beta^0$, $\beta_s^1 = 1$	1	0.759	0.557	0.517	0.517	0.517	0.515
VA2	Eqn. (27) with $\beta_s^1 = 1$	0.759	1	0.557	0.517	0.517	0.518	0.516
L1	Eqn. (27) with $\beta_s^0 = \beta^0$, $\beta_s^1 = \beta^1$	0.557	0.557	1	0.863	0.863	0.862	0.863
L2	Eqn. (27) with $\beta_s^1 = \beta^1$	0.517	0.517	0.863	1	0.877	0.875	0.877
L3	Eqn. (27)	0.517	0.517	0.863	0.877	1	0.877	0.878
DUM	Eqn. (4) with $\eta = \infty$	0.517	0.518	0.862	0.875	0.877	1	0.88
SEM	Eqn. (4) with $\eta = 2 \times 10^{-5}$	0.515	0.516	0.863	0.877	0.878	0.88	1

As shown in Table 9, value-added models (**VA1** and **VA2**) can produce assessments of school performance that are quite different from those obtained from the linear, dummy variable or semiparametric models. Specifically, the correlation between rankings from the value-added and smoothed semiparametric models was approximately .52. *Given our results in section 4.3, we conclude that differences in school performance assessments between the preferred semiparametric model and the value-added specification are greater than differences produced when adding student-level demographic controls to a particular model.* This result is particularly striking since many accountability policies and studies in economics

have stressed the need to control for demographic characteristics, yet the specification of the relationships between intake and follow-up scores has largely been overlooked. Our analysis suggests that these specification issues are of first-order importance, particularly when considering the value-added model against alternate specifications.

We explain the low correlation between value-added rankings and the rankings obtained in our other models by noting that the value-added specification fails to capture the shape of the relationships between intake and expected follow-up achievement. In particular, the graphs in Figure 1 strongly show that initially low achieving students tend to demonstrate the most improvement, as the gaps between the regression functions and the 45 degree lines are largest at the left-tail of the intake score distribution. Value-added models will thus tend to rank schools with initially low-achieving students highly.[31] However, this will not be the case for the semiparametric (and to a lesser extent, linear) models, since these models first estimate expected performance profiles and then rank school performance off of these estimated relationships. These results seem to call into question the use of value-added specifications for performance assessment and policy evaluation.

5 Conclusion

In this paper we described a new and flexible framework for modeling school effects using Bayesian hierarchical models with smoothing priors. Our "smoothed" analysis provided reasonable depictions of the relationships between intake and expected follow-up achievement, and statistically our model was found to be favored over a variety of competitors. These competing specifications include the popular value-added and linear models that form the backbone for previous research and underlie numerous school accountability policies.

Using data from High School and Beyond (HSB) we found that rankings of school performance were sensitive to the models employed, and in particular, value-added specifications could give widely different and potentially misleading performance assessments. The failure of the value-added specification arises since the use of gain scores rewards improvements equally across the intake achievement support, and does not recognize that low-achieving students (and schools) tend to demonstrate the most improvement on average. Finally, when estimating the popular value-added model in a generalized hierarchical setting, we found little evidence that proxies for school quality played any role in explaining variation in performance across schools, as these variables would not be reported as "significant." However, when estimating our preferred semiparametric specification, we found small but reasonably significant impacts of school quality proxies such as district-level expenditure per pupil.

[31] As suggestive evidence of this, note that the average intake scores for the top 10 schools as ranked in value-added specification were .142, .192, .309, .338, .357, .366, .371, .381, .382 and .386, respectively. An identical pattern emerges when calculating the average intake scores by quartile and decile of the school rankings. These results strongly suggest an inverse relationship between intake achievement and value-added rankings.

Appendix: The Gibbs Sampler

Standard arguments show that the model outlined in (4) - (6) together with the priors

$$\beta_s|\eta_s,\sigma^2_{ys} \overset{ind}{\sim} N\left(0,\sigma^2_{ys}H_s^{-1}V(\eta_s)[H_s^{-1}]'\right) \equiv N(0,\sigma^2_{ys}\Omega_s), s=1,2,\cdots,\overline{S} \tag{34}$$

$$\sigma^2_{ys} \overset{ind}{\sim} IG(\underline{e}_{1s},\underline{e}_{2s}), \quad s=1,2,\cdots\overline{S} \tag{35}$$

$$\sigma^2_{\alpha} \sim IG(\underline{a}_1,\underline{a}_2) \tag{36}$$

$$\sigma^2_{\gamma} \sim IG(\underline{g}_1,\underline{g}_2) \tag{37}$$

yield the joint posterior distribution $p(\theta|\text{Data})$ up to proportionality. In the above we have defined $\Omega_s = \Omega_s(\eta_s) = H_s^{-1}V(\eta_s)[H_s^{-1}]'$ and we let θ denote all the model parameters, i.e.,

$$\theta = [\{\alpha_{ih}\}_{i=1}^{N}\ \{\gamma_h\}_{h=1}^{H}\ \{\beta_s\}_{s=1}^{\overline{S}}\ \{\sigma^2_{ys}\}_{s=1}^{\overline{S}}\ \sigma^2_{\alpha}\ \sigma^2_{\gamma}].$$

We implement the Gibbs sampler (e.g., Casella and George (1992)) to fit this model, which involves successively sampling from the posterior conditionals. To mitigate the degree of autocorrelation in our parameter chains, we introduce an algorithm which samples the slope coefficients $\{\beta_s\}$, individual effects $\{\alpha_{ih}\}$ and school effects $\{\gamma_h\}$ in a *single block*. We do this by first sampling from the conditional for the regression parameters $\{\beta_s\}$ which is *marginalized over the individual and school effects*. We denote this conditional as $p(\beta|\theta_{-\beta,\alpha,\gamma},\text{Data})$, where $\beta \equiv \{\beta_s\}$ and θ_{-x} denotes all parameters in the model other than x. We then sample from the conditional posterior distribution of the individual random effects *marginalized over the school effects:* $p(\alpha|\theta_{-\alpha,\gamma},\text{Data})$. Finally, the blocking step is completed by sampling from the complete posterior conditional for the school random effects $p(\gamma|\theta_{-\gamma},\text{Data})$.

We will let $\overline{S}$ denote the total number of subject tests s, H denote the total number of (high) schools in the sample, n_h denote the number of the students in school h, and $\overline{J} = \sum_{s=1}^{\overline{S}} J_s$. Finally, define β as the $\overline{J}\times 1$ vector

$$\beta = [\beta_1'\ \beta_2'\ \cdots \beta_{\overline{S}}']'.$$

To derive the conditionals used in the sampler we will need to stack observations to the school level and define some additional notation. We can write equation (4) as

$$y_h = D_h\beta + \epsilon_h + u_h + \iota_{n_h\overline{S}}v_h,$$

where $\iota_{n_h\overline{S}}$ denotes an $(n_h\overline{S})\times 1$ vector of ones and we have rewritten (6) as $\gamma_h = 0 + v_h$, where $v_h \overset{iid}{\sim} N(0,\sigma^2_{\gamma})$. The variables y_h, D_h, ϵ_h, and u_h are quantities stacked to the school level, defined as follows:

$$y_h = \begin{bmatrix} y_{1h} \\ y_{2h} \\ \vdots \\ y_{n_hh} \end{bmatrix}, \quad \text{where} \quad y_{jh} = \begin{bmatrix} y_{j1h} \\ y_{j2h} \\ \vdots \\ y_{j\overline{S}h} \end{bmatrix}.$$

$$D_h = \begin{bmatrix} D_{1h} \\ D_{2h} \\ \vdots \\ D_{n_h h} \end{bmatrix}, \quad \text{where} \quad D_{jh} = \begin{bmatrix} D_{j1h} & 0 & \cdots & 0 \\ 0 & D_{j2h} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & D_{j\overline{S}h} \end{bmatrix}.$$

$$\epsilon_h = \begin{bmatrix} \epsilon_{1h} \\ \epsilon_{2h} \\ \vdots \\ \epsilon_{n_h h} \end{bmatrix}, \quad \text{where} \quad \epsilon_{jh} = \begin{bmatrix} \epsilon_{j1h} \\ \epsilon_{j2h} \\ \vdots \\ \epsilon_{j\overline{S}h} \end{bmatrix}.$$

Finally,

$$u_h = \begin{bmatrix} \iota_{\overline{S}} u_{1h} \\ \iota_{\overline{S}} u_{2h} \\ \vdots \\ \iota_{\overline{S}} u_{n_h h} \end{bmatrix}.$$

Given the Normality assumptions in (4) - (6), it follows that

$$y_h|\beta, \Sigma, \sigma_\alpha^2, \sigma_\gamma^2 \sim N\left(D_h\beta, \left[I_{n_h} \otimes \left[\Sigma + \sigma_\alpha^2 \iota_{\overline{S}} \iota_{\overline{S}}'\right] + \sigma_\gamma^2 \iota_{n_h \overline{S}} \iota_{n_h \overline{S}}'\right]\right),$$

where

$$\Sigma \equiv \begin{bmatrix} \sigma_{y1}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{y2}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{y\overline{S}}^2 \end{bmatrix}.$$

Because of the assumed independence of observations across schools, we obtain

$$\beta|\theta_{-\beta,\alpha,\gamma}, \text{Data} \sim N(D_\beta d_\beta, D_\beta),$$

where

$$D_\beta = \left[\sum_{h=1}^{H} \left[D_h' \left(I_{n_h} \otimes \left[\Sigma + \sigma_\alpha^2 \iota_{\overline{S}} \iota_{\overline{S}}'\right] + \sigma_\gamma^2 \iota_{n_h \overline{S}} \iota_{n_h \overline{S}}'\right)^{-1} D_h\right] + \Omega^{-1}\right]^{-1},$$

$$d_\beta = \sum_{h=1}^{H} D_h' \left(I_{n_h} \otimes \left[\Sigma + \sigma_\alpha^2 \iota_{\overline{S}} \iota_{\overline{S}}'\right] + \sigma_\gamma^2 \iota_{n_h \overline{S}} \iota_{n_h \overline{S}}'\right)^{-1} y_h,$$

and $\Omega = \Omega\left(\{\sigma_{ys}^2, \eta_s\}_{s=1}^{\overline{S}}\right) \equiv diag\{\sigma_{ys}^2 \Omega_s\}_{s=1}^{\overline{S}}$. For the posterior conditional for the set of individual effects α_{ih} marginalized over the school effects γ_h, let $\alpha = [\alpha_{1h}\ \alpha_{2h}\ \cdots\ \alpha_{Nh}]$ and note that the independence across schools assumed in the model implies

$$p(\alpha|\theta_{-\alpha,\gamma}, \text{Data}) = \prod_{h=1}^{H} p(\alpha_h|\theta_{-\alpha_h,\gamma}, \text{Data}),$$

where $\alpha_h = [\alpha_{1h}\alpha_{2h}\cdots\ \alpha_{n_h h}]'$ denotes the vector of individual random effects from school h. This implies that a draw from the desired conditional for α can be obtained by drawing independently from the α_h conditionals. Integrating γ_h out of the prior for α_{ih} in (5), and stacking observations in (4) to the school level, we obtain

$$\alpha_h|\theta_{-\alpha_h,\gamma}, \text{Data} \overset{ind}{\sim} N(D_{\alpha_h} d_{\alpha_h}, D_{\alpha_h}), \quad h = 1, 2, \cdots, H$$

where

$$D_{\alpha_h} = \left[\sum_{s=1}^{\overline{S}} \sigma_{ys}^{-2} I_{n_h} + \left[\sigma_\alpha^2 I_{n_h} + \sigma_\gamma^2 \iota_{n_h}\iota_{n_h}'\right]^{-1}\right]^{-1},$$

$$d_{\alpha_h} = \overline{S} \begin{bmatrix} \overline{y}_{1h} \\ \overline{y}_{2h} \\ \vdots \\ \overline{y}_{n_h h} \end{bmatrix}$$

and

$$\overline{y}_{jh} = \frac{1}{\overline{S}} \sum_{s=1}^{\overline{S}} \frac{y_{jsh} - D_{jsh}\beta_s}{\sigma_{ys}^2}.$$

Finally, the blocking step is completed by drawing the school effects independently from their complete posterior conditional

$$\gamma_h|\theta_{-\gamma_h}, \text{Data} \overset{ind}{\sim} N\left(\left[n_h/\sigma_\alpha^2 + \sigma_\gamma^{-2}\right]^{-1} \sum_{i=1}^{n_h} \alpha_{ih}/\sigma_\alpha^2, \left[n_h/\sigma_\alpha^2 + \sigma_\gamma^{-2}\right]^{-1}\right), \quad h = 1, 2, \cdots, H.$$

The remaining variance parameters are drawn from their complete posterior conditional distributions. For the first-stage variance parameters, we obtain (for $s = 1, 2, \cdots, \overline{S}$):

$$\sigma_{ys}^2|\theta_{-\sigma_{ys}^2}, \text{Data} \overset{ind}{\sim} IG\left(\frac{N + J_s}{2} + \underline{e}_{1s}, \left[\underline{e}_{2s}^{-1} + \frac{1}{2}\sum_{i=1}^{N}(y_{ish} - \alpha_{ih} - D_{ish}\beta_s)^2 + \frac{1}{2}\beta_s'\Omega_s^{-1}\beta_s\right]^{-1}\right).$$

Finally, we obtain the complete conditionals for the individual and school level variance parameters:

$$\sigma_\alpha^2|\theta_{-\sigma_\alpha^2}, \text{Data} \sim IG\left(\frac{N}{2} + \underline{a}_1, \left[\underline{a}_2^{-1} + \frac{1}{2}\sum_{i=1}^{N}(\alpha_{ih} - \gamma_h)^2\right]^{-1}\right)$$

and

$$\sigma_\gamma^2|\theta_{-\sigma_\gamma^2}, \text{Data} \sim IG\left(\frac{H}{2} + \underline{g}_1, \left[\underline{g}_2^{-1} + \frac{1}{2}\sum_{h=1}^{H}\gamma_h^2\right]^{-1}\right).$$

References

Aitkin, M. and N. Longford (1986), "Statistical Modeling Issues in School Effectiveness Studies (with discussion)," *Journal of the Royal Statistical Society, Series A*, 149, 1-42.

Betts, J. (1995), "Does School Quality Matter? Evidence from the National Longitudinal Survey of Youth," *Review of Economics and Statistics*, 77(2), 231-250.

Carlin, B.P. and T. Louis (2000), *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd edition, Boca Raton: Chapman & Hall.

Casella, G. and E. George (1992), "Explaining the Gibbs Sampler," *The American Statistician*, 46, 167-174.

Goldstein, H. and D.J. Spiegelhalter (1996), "League Tables and Their Limitations: Statistical Issues in Comparisons of Institutional Performance," *Journal of the Royal Statistical Society, Series A*, 159(3), 385-443.

Goldstein, H. and S. Thomas (1996), "Using Examination Results as Indicators of School and College Performance," *Journal of the Royal Statistical Society, Series A*, 159(1), 149-163.

Hanushek, E., J. Kain and S. Rivkin (1998), "Teachers, Schools, and Academic Achievement," *NBER Working Paper No. 6691.*

Hoffer, T., A.M. Greeley and J.S. Coleman (1985), "Achievement Growth in Public and Catholic Schools," *Sociology of Education*, 58(2), 74-97.

Kane, T.J. and D.O. Staiger (2002), "The Promise and Pitfalls of Using Imprecise School Accountability Measures," *Journal of Economic Perspectives*, 16(4), 91-114.

Koop, G. and D. J. Poirier (2004a), "Bayesian Variants of Some Classical Semiparametric Regression Techniques," *Journal of Econometrics*, 123(2), 259-282.

Koop, G. and D. J. Poirier (2004b), "Empirical Bayesian Inference in a Nonparametric Regression Model," to appear in a volume from the *Conference in Honour of Professor J. Durbin on State Space Models and Unobserved Components.*

Lee, V.M. and J.B. Smith (1995), "Effects of High School Restructuring and Size on Early Gains in Achievement and Engagement," *Sociology of Education*, 68(4), 241-270.

Lindley, D. and A.F.M. Smith (1972), "Bayes Estimates for the Linear Model," *Journal of the Royal Statistical Society, Series B*, 34, 1-41.

Link, C.R. and J.G. Mulligan (1991), "Classmates' effects on black student achievement in public school classrooms," *Economics of Education Review*, 5(4): 297-310.

Poirier, D.J. (1995), *Intermediate Statistics and Econometrics*, Cambridge: The MIT Press.

Raftery, A.E. (1996), "Hypothesis Testing and Model Selection," in *Markov Chain Monte Carlo in Practice*, W.R. Gilks, S. Richardson and D.J. Spiegelhalter, eds., Boca Raton: Chapman & Hall, 163-187.

Willms, J.D. (1985), "Catholic-School Effects on Academic Achievement: New Evidence from the High School and Beyond Follow-up Study," *Sociology of Education*, 58(2), 98-114.

Yang, M., H. Goldstein, W. Browne and G. Woodhouse (2002), "Multivariate Multilevel Analyses of Examination Results," *Journal of the Royal Statistical Society, Series A*, 165(1), 137-153.