IOWA STATE UNIVERSITY

Schools, School Quality and Academic Achievement: Evidence from the Philippines

Marigee Bacolod, Justin Tobias

February 2005

Working Paper # 05006

Department of Economics
Working Papers Series

Ames, Iowa 50011

Schools, School Quality and Achievement Growth: Evidence From the Philippines

September 9, 2004

Marigee Bacolod
mbacolod@uci.edu
University of California-Irvine
3151 Social Science Plaza A
Irvine, CA 92697-5100

Justin L. Tobias
tobiasj@iastate.edu
Iowa State University
260 Heady Hall
Ames, IA 50011-1070

Abstract

A broad literature seeks to assess the importance of schools, proxies for school quality, and family background on children's achievement growth using the education production function. Using rich data from the Philippines, we introduce and estimate a model that imposes little structure on the relationship between intake achievement and follow-up achievement and evaluate school performance based on this estimated relationship. Our methods nest typical value added specifications that use test score gains as the outcome variable and models assuming linearity in the relationship between intake and follow-up scores. We find evidence against the use of value-added models for our data and show that such models give very different assessments of school performance in the Philippines. Using a variety of tests we find that schools matter in the production of student achievement, though variation in performance across schools only explain about 6 percent of the total (conditional) variation in follow-up achievement. Schools providing basic facilities - in particular schools providing electricity - are found to perform much better in the production of achievement growth.

---

1 Introduction

The current debate on school quality can arguably trace its roots to the publication of the Coleman Report (Coleman et. al. 1966), which claimed that family background and the characteristics of other students in U.S. schools were more important than school differences in accounting for variation in educational outcomes. In detailed reviews of subsequent work in the U.S. and in developing countries, Hanushek (1986, 1995, 2002) concludes that there is little systematic relationship between various educational inputs and student performance. Despite this predominant view, a handful of studies have found that teachers and schools matter in terms of student achievement, e.g., Hanushek, Kain, and Rivkin (1998), Ehrenberg and Brewer (1994).

A long-running staple in the literature for identifying the relative importance of measurable educational inputs is the education production function. Analogous to firm production, this framework relates contemporaneous child cognitive outcomes with the entire history of family inputs, school inputs and innate ability. When performances on comparable test instruments are available over time, work on this topic has adopted one of two estimation approaches. The "value-added" approach assesses school performance by modeling changes in test scores (test score gains) as the dependent variable, thus differencing out time-invariant unobservable factors (*e.g.* Hanushek et. al. 1998). An alternate approach writes the follow-up score as a linear function of the baseline score (*e.g.* Link and Mulligan 1991). Both approaches share common goals of empirically evaluating the importance of schools, determining the fraction of outcome variation that is attributable to schools, and finding school characteristics that can explain variation in school performance.

In this paper we offer two primary contributions. First, within the framework of observational studies with student-level test scores over time (that we denote as an intake/baseline score and a follow-up score), we introduce and estimate a model that generalizes those previously employed in the literature. This model nests traditional value-added specifications and those models assuming linearity in the relationship between intake and follow-up scores. Formally, the model we estimate is a *semiparametric hierarchical regression model*. Though the econometrics involved are reasonably detailed, the intuition behind our approach is straightforward and also reveals how our method for assessing school performance differs from methods assumed in traditional gain-score models.

For each student in our data with a given intake score, we obtain a predicted follow-up score. These predicted scores are obtained by *locally* averaging the follow-up scores of students *across all schools* who received similar intake scores. If an individual student's observed follow-up score lies above her predicted score, then she has improved more than other students with similar intake achievement. We then collect these student level "residuals" - differences between observed follow-up scores and predicted scores - and assign them into groups according to school attended. School rankings and

estimates of school effects are then obtained by averaging the residuals within schools and comparing the averages across schools. If the average for a given school is high, then many of the students in that school have improved more than students at other schools with comparable baseline achievement. We use this method to form estimates of school performance and then search for school-level observables that explain variation in these school effects.

The extension of a standard hierarchical model to one with a semiparametric component is not just a statistical contribution, but its use proves to be appropriate for this application. Beyond this, the model we describe and the algorithm we derive for fitting it should be useful for other studies in this area. To see the potential usefulness of our approach, it is important to recognize that popular value-added studies reward achievement growth equally across the level of intake achievement. That is, when using score gains as the dependent variable, no distinction is made between, say, an average improvement of 10 points for an initially high achieving school and a 10 point improvement for an initially low achieving school. We will show in this paper that such comparisons should not be made across the initial achievement distribution, as it is much more difficult for initially high achieving schools (and students) to match the level improvements of initially low achieving schools (and students). As such, we first predict how much students should improve based on their intake achievement, and then evaluate the performance of schools from this flexibly estimated relationship. We compare our model's assessments of Philippine schools with those from linear and value-added models, and find strong and convincing evidence against the use of value-added specifications for our data.

The second contribution of this paper is our investigation of the roles of schools and proxies for school quality, in this case applied to the Philippines. Because most observational data, particularly those from developing countries, do not contain more than one achievement score per student, most previous work in this area has regressed contemporaneous outcomes on contemporaneous inputs. As such, these studies might be fraught with omitted variables bias.[1] To mitigate these concerns, these nonexperimental studies have typically controlled for a wide set of measures of resources and process factors, including pedagogical process, teacher preparation and school organization. The idea behind the "kitchen-sink" approach is that one can form a more complete picture of school effects and what policies matter by adding more measures to explain students' achievement levels.

The methods we employ in this paper differ from the approaches used in many U.S. and developing country studies in that we have data on student achievement at two different points in time. We thus

[1]The education production function captures the theoretical notion that child cognitive achievement is a cumulative process depending on a history of inputs. When data are available for at least two years, the lagged test score can approximate the cumulative contribution of all historical inputs and innate ability. Studies with contemporaneous information tend to control for school choice selectivity to mitigate the bias. See for example, the set of studies reviewed in Hanushek (1995, 2002) and Glewwe (2002). Glewwe also makes the point that most work in developing countries has been based on survey data where tests are administered only once.

evaluate schools and the empirical importance of proxies for school quality by using student-level *improvements* in test scores as the evaluation metric rather than cross-sectional differences in level scores. In the case of non-random assignment of students to schools on the basis of achievement, studies looking at only cross-sectional variation in test scores could confound school effects with underlying characteristics of the students attending those schools. In our study, we control for intake achievement and a variety of other student-level demographic information directly, and thus mitigate most of the concerns regarding selection and omitted variables biases in observational studies, as noted by Hanushek (1986).

In addition, our study differs from most previous work in developing countries as we first establish the existence of school effects and then seek to explain why it is that schools differ. This distinction is important because previous studies of measured resources - which typically look at only cross-sectional variation across schools - do not first provide clear evidence of the existence of underlying differences among schools. In his review of 96 previous papers in developing countries, Hanushek (1995, p.234) supports this notion as he states that previous studies "do not identify the overall systematic variations in school quality." [2] The approach employed in this paper enables us to test if school effects are present, to find characteristics that explain variation in school performance, and to quantify the extent of variation across schools that remains unobserved.

We find compelling evidence that schools in the Philippines differ and matter in the production of achievement and implement a variety of intuitive tests to document the existence of these school effects. While evidence regarding the impact of teacher and school characteristics on academic achievement in developed countries is mixed, studies in developing countries generally provide evidence that certain basic school resources matter. For example, Harbison and Hanushek (1992) find that although the pupil-teacher ratio and teacher characteristics had inconsistent effects among Brazilian school children, indexes of building facilities and writing materials had significantly positive impacts across various specifications. Fuller and Clark (1994), in a review of work through the mid-nineties, find that only 9 out of 26 primary-school studies and only 2 of 22 secondary-school studies show a significant impact of class size on student achievement in developing countries. These results suggest that basic resources matter in the production of achievement in a developing country, while more traditional proxies for school quality have little impact.

Our results using data from the Philippines generally confirm these findings. Minimal basic facilities, and in particular, the provision of electricity, matter more than class-size and teacher training programs. The traditional school-level controls we use, which include class size, teacher training, teacher experience, and an electricity indicator, help to explain away a substantial portion of the variation in performance across schools. Finally, although schools clearly differ and are important

[2] More recent work, including Duflo (2001), tend to exploit interesting policy or natural experiments and identify the effect of specific policy interventions.

determinants of student achievement in the Philippines, we find that schools account for only 6 percent of the total variation in student-level follow up scores.

The paper is organized as follows. Section 2 describes our model and briefly describes our approach to estimation. Section 3 describes our data and empirical results are presented in section 4. The paper concludes with a summary in section 5. Figures, estimation details and descriptions of our data are provided in appendices A-C, respectively.

2 The Model

The most general model we use to evaluate schools and the empirical importance of school characteristics in the Philippines can be expressed as follows:

$$y_{is} = \alpha_s + f(x_{is}) + z_{is}\beta + \epsilon_{is}, \qquad \epsilon_{is} \overset{iid}{\sim} N(0, \sigma^2_\epsilon) \tag{1}$$

$$\alpha_s = Q_s\gamma + u_s, \qquad u_s \overset{iid}{\sim} N(0, \sigma^2_\alpha). \tag{2}$$

In the above, y_{is} refers to the follow-up test for individual i attending school s, α_s is a school-level effect, x_{is} denotes the baseline test score, z_{is} denotes a vector of additional student-level controls, and Q_s denotes a vector of school quality variables. The quantity σ^2_α can be interpreted as the amount of variation in school performance that remains unobserved after controlling for characteristics Q. By estimating models with and without Q, we can determine how much of the variation in school performance can be explained through observable school characteristics.

There are a number of features of this model that merit discussion. First, and perhaps most importantly, the model permits a very flexible conditional mean specification that relates the baseline test score to the follow-up test score through the function f. In particular we allow for the possibility of nonlinearities in this relationship, and thus can test if often-used parametric specifications are appropriate. For example, "value added" specifications use gain scores as the dependent variable, thus imposing $f(x) = \delta_0 + x$, while other studies regress the follow-up test on the baseline test, implicitly imposing $f(x) = \delta_0 + \delta_1 x$.

Our flexible modeling strategy will enable us to determine if these assumptions are appropriate for modeling student achievement in the Philippines, and may also be suggestive of the appropriateness of such assumptions in other data sets. Intuitively, we might expect initially low achieving students to show a relatively large improvement in test scores, while initially high achieving students may only be expected to maintain their high achieving status. This intuition suggests that one should be careful to let the data reveal the shape of the relationship between baseline scores and follow-up scores rather than bring a potentially inappropriate specification to the data.

School effects are captured in the α_s parameters. These parameters *jointly* capture data information from school-specific outcomes in (1) and information that expresses some degree of commonality across the school effects in (2). The α_s parameters can be interpreted as a type of school-level "residual." For every individual in the sample we obtain a predicted follow-up score given an initial test score. Since our treatment of this relationship is nonparametric, predicted follow-up scores can be interpreted as local averages of follow-up scores for students *across schools* who received similar initial test scores. Students whose observed follow-up scores are higher than their predicted values have improved more than other students who received similar initial scores. School effects are then defined as averages of these individual "residual" values. Schools with high average "residuals" contain many students who have improved more than other students with similar intake achievement, and we would label such schools as high-achieving. Allowing for a nonparametric specification through f enables us to correct for average differences in achievement growth across the initial achievement distribution, and then to evaluate school performance based on this estimated relationship.

In equation (2) we also introduce school-level characteristics that may potentially explain variation in performance across schools. In particular, equation (2) centers the mean of the school-level effects around measures of the "quality" of the school, $Q_s\gamma$.[3] We will investigate if often-cited proxies for school quality such as average class sizes and average teacher training can explain the observed variation in performance across schools. Given that our data is from the Philippines we are also able to investigate if basic facilities associated with the school - such as the availability of electricity - play any role in school performance.

2.1 Estimation

The model in (1) - (2) requires an investment in some econometric methodology, as it requires joint estimation of school-specific random effects (α_s) together with the *function* f. In this paper we adopt a simulation-based Bayesian approach for estimating the model described by (1) - (2), and find this approach to be attractive for a variety of reasons. First, use of this approach enables us to obtain exact finite-sample inference for the school level parameters α_s. This is particularly important for us since we do not observe a large number of observations per school for all schools, and thus would not want to rely exclusively on asymptotic results to make inference regarding these parameters. Second, the simulation-based algorithm we derive and apply (fully described in Appendix B) is very well suited for handling estimation of the nonparametric function f simultaneously with the school-level random effect parameters α_s. In particular, the algorithm we employ (the Gibbs sampler) enables us to break the joint estimation problem into an equivalent series of smaller and more manageable problems. Finally, our particular method for nonparametric estimation is straightforward and

[3]For identification purposes, we do not include an intercept in Q (as it is absorbed in the constant when estimating f) and also standardize each element of Q to have mean zero.

intuitive and is not dependent on choices of bandwidths or smoothing parameters.

Estimation of the Nonparametric Component

To be sure, there are a number of ways to approach nonparametric estimation of regression functions. Some popular alternatives include kernel methods (*e.g.* Fan and Gijbels (1996), DiNardo and Tobias (2001)), methods involving smoothness priors (*e.g.* Koop and Poirier (2003), Chib and Jeliazkov (2003)), and splines and roughness penalty approaches (Green and Silverman (1994)) . In this paper, we choose among these alternatives and estimate the function f using a cubic regression spline:

$$f(x) = \delta_0 + \delta_1 x + \delta_2 x^2 + \sum_{j=3}^{J} \delta_j (x - \tau_{j-2})_+^3,$$

where

$$z_+ \equiv \max(0, z).$$

What separates our analysis from that of a standard spline analysis is our treatment of the $J-2$ knot points $\{\tau_j\}_{j=1}^{J-2}$. These knots are added to the specification of f to flexibly permit changes in the curvature of the regression function. Unlike many studies that declare the number and location of the knots *a priori*, we treat knot location and quantity as parameters to be determined within our model. Specifically, our estimation approach is quite similar to that of Smith and Kohn (1996) who introduce a relatively fine grid of *potential* knots, and then let the data decide which knots to keep. As with parameter estimation, there will be uncertainty in the estimation of our function through uncertainty regarding the number and location of the knot points. As such, our algorithm will take into account this source uncertainty in constructing confidence intervals for the estimated regression function. The algorithm is also *adaptable* in the sense that if the true model is, say, linear, the posterior will place little probability on keeping any of the potential knots, and will consequently restrict our flexible spline function to the linear model.

We performed numerous generated data experiments to ensure that our code was error-free and that our method provided accurate estimates of f, $\{\alpha_s\}$, σ_α^2, σ_ϵ^2 and β even when the true f was quite nonlinear. MATLAB code and the data for fitting this model are available upon request, and specific details regarding our algorithm can be found in Appendix B. Of course, our Bayesian approach requires us to choose priors for our model parameters, and the specifics of our prior specification are provided in Appendix B. We select our priors to be non-informative or diffuse (yet computationally convenient) so that the data information is vastly predominant. In sensitivity analyses, substantive conclusions and parameter point estimates were not significantly affected by moderate changes in our prior specification.

3 The Data

We use data from the Cebu Longitudinal Health and Nutrition Survey (CLHNS), which was carried out in the Metropolitan Cebu area on the island of Cebu, Philippines. Metro Cebu includes Cebu City, the second largest city in the Philippines, and several surrounding urban and rural communities. The CLHNS tracks a sample of 3,080 children born between May 1, 1983 and April 30, 1984, in randomly selected barangays (districts). In 1991-92 and 1994-95, follow-up surveys of mothers and children were conducted, and IQ tests were administered to the children. In 1994-95 and also in 1996-97, English reading comprehension and mathematics tests were developed for the surveys based on official school curricula at various grades. The tests were administered to the index children (that is, the children surveyed starting in the first round) and to their younger sibling of schooling age. These follow-up surveys collected detailed schooling history of each index child and, if in school, his or her younger sibling.[4] The 1994-95 follow-up surveys also gathered detailed information on the schools the children attended, including academic inputs and teacher characteristics.

In this paper, we focus on achievement in mathematics and thus use the 1994-95 mathematics test score as our measure of intake achievement, and use the 1996-1997 mathematics test score as the measure of follow-up achievement. Tests were carefully designed to be comparable over time, and in each case a value of 60 corresponded to a perfect score.

Excluding children with missing test scores, children who were not enrolled in the 1996-97 school year, and children who transferred schools between 1994-95 and 1996-97 leaves us with 2,136 children for our analysis. [5] Means of key variables used in the analysis are reported in Table 1. As the sample is later restricted by identifying schools that were attended by three or more, five or more, and ten or more children in the sample, means are further broken down along these lines.

[4] A third follow-up survey (1998-99) is being processed. As the 1998-99 surveys occurred when most adolescents were already in their second or third year in high school, our study does not include the latest survey round.

[5] Sample selection and attrition are further tabulated in Appendix C.

Table 1
Descriptive Statistics of Key Variables

	Sample Restriction							
	ALL ($n = 2136$)		$n_s \geq 3$ ($n = 2047$)		$n_s \geq 5$ ($n = 1976$)		$n_s \geq 10$ ($n = 1801$)	
VARIABLES	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
Test94	14.78	(13.81)	14.46	(13.66)	14.32	(13.63)	14.06	(13.37)
Test96	23.16	(14.90)	22.90	(14.79)	22.83	(14.77)	22.50	(14.62)
Male	0.515	(0.50)	0.514	(0.50)	0.513	(0.50)	0.51	(0.50)
DadEd	6.55	(4.41)	6.55	(4.37)	6.52	(4.32)	6.44	(4.23)
MomEd	6.74	(4.02)	6.68	(3.95)	6.63	(3.91)	6.54	(3.78)
LogExpend	7.28	(1.34)	7.25	(1.32)	7.23	(1.31)	7.19	(1.29)
Characteristics of Schools Attended in 1996-97								
ClassSize	43.53	(13.53)	44.80	(11.45)	44.45	(11.28)	43.62	(9.88)
TeachExp	13.5	(7.29)	14.80	(7.05)	14.55	(7.36)	15.24	(7.29)
InService	86.92	(20.49)	88.55	(18.43)	88.70	(16.21)	88.70	(13.94)
Electricity	0.83	(0.38)	0.83	(0.38)	0.81	(0.40)	0.77	(0.43)
Private	0.20	(0.40)	0.17	(0.38)	0.14	(0.35)	0.09	(0.28)
BookStu	2.36	(5.74)	1.63	(3.19)	1.27	(2.29)	1.39	(2.65)
FacIndx	7.80	(1.98)	7.84	(1.94)	7.82	(1.86)	7.68	(1.63)

There are three additional features of our data that are worth noting. First, even when we restrict our data to include only schools with ten or more children in the sample, we still have 47 schools in our sample with an average of 38 students per school. When relaxing the requirement of at least 10 observations per school, we gain additional schools, but also reduce the average number of observations per school. Given the potential for coefficient estimates to change as we vary our required number of observations per school, we choose to estimate our models across the three sample restrictions ($n_s \geq 3, 5, 10$) to investigate the robustness of our results.

Second, although family income is often cited as one of the most important family background characteristics to control for, it is difficult to measure because of the pervasiveness of self-employment in the Philippines. Appealing to previous developing country studies, we instead use total household expenditures divided by the number of family members as our family resource measure. We also control for mother's and father's years of schooling in addition to log per-capita household expenditures. Note that less than half of the parents of children in our samples attained more than an elementary education.

Finally, the school questionnaires collected detailed information on many primary school characteristics. In our analysis we focus on frequently used proxies for school quality, such as class size and average teacher training, as well as basic facilities. Class sizes–defined as the total number of students in the school divided by the number of sessions taught–tend to be very large in Cebu schools, averaging 44 to 45 students per class. Teachers also tend to be older, with average teaching

experience (TeachExp) of 14 to 15 years. The vast majority of the teachers in our sample have had in-service training (InService), averaging at 87 to 89 percent across schools.

Importantly for our purposes, not all schools in Cebu have basic facilities. The variable FacIndx sums over the following indicator variables indicating the quality of the school buildings: whether concrete is the main construction material, if the school has electricity, if water supply is piped in, if the school has flush or water-sealed toilets (as opposed to open pit or latrines), if the school has 100 percent usable blackboards, whether the school has a library, and direct indicators of classroom environment: no classes held outside due to lack of space, no multigraded classrooms, no temporary partitions in classrooms, and no classes have to be moved or excused when it rains. Schools that the children in our samples attended had on average 7.7 to 7.8 out of 10 of these facilities. School libraries also have little more than 1.3 to 2.4 books per child on average (BookStu). [6] As many as 17 to 23 percent of schools in our sample are without any electricity. As pointed out earlier, an emerging theme among studies in developing countries is that the provision of basic resources matter while teacher characteristics and class sizes have little effect on student achievement. We will explore the effect of all these various school quality measures in explaining school-level performance variation in the Philippines.

4 Empirical Results

We are interested in employing the model described in (1) -(2) and variants of it to address the following key questions using the CLHNS: (1) Is the relationship between intake and follow-up scores nonlinear? (2) If so, what are the implications of this result for value-added specifications and models assuming linearity in assessments of school performance? (3) Do schools matter in the production of student achievement in the Philippines? (4) If so, how much variation in student outcomes is attributable to variation in performance across schools? (5) Can we find characteristics of schools that are associated with good performance? In particular, are class sizes, teacher training, and the provision of basic facilities systematically related with school performance? (6) Do other individual controls like family characteristics and parental resources also affect follow-up achievement even after controlling for intake achievement?

To begin to explore the relationship between baseline and follow-up test scores and to examine the empirical importance of school effects in determining student achievement, we estimate a simplified

[6] and with strictly positive library holdings, the mean number of books per student ranges from 1.6 to 3.

version of our model in (1) and (2):[7]

$$y_{is} = \alpha_s + f(x_{is}) + \epsilon_{is}, \qquad \epsilon_{is} \overset{iid}{\sim} N(0, \sigma^2_\epsilon) \tag{3}$$

$$\alpha_s \overset{iid}{\sim} N(0, \sigma^2_\alpha) \tag{4}$$

We begin with this restricted model since we first want to determine if school effects are present before to trying to find covariates Q that explain variation in performance at the school level. By excluding Q from the second stage, we are also able to quantify the overall degree of of performance variation at the school level, σ^2_α, and to determine the fraction of total variation in follow-up outcomes that is attributable to variation across schools. Additionally, we do not, as yet, add any covariates z_{is} to equation (1) to fix ideas on the nature of the relationship between baseline and follow-up scores. We will incorporate both sets of covariates in later elaborations of this model.

4.1 The Shape of the Regression Function

In Figure 1 we present our point estimates and standard error bands associated with the function $f(x)$, which represents the conditional expectation of follow-up t estimated regression function tends to flatten out. The estimated function is increasing throughout the support of initial test scores, indicating that higher baseline scores lead to higher expected follow-up scores, though the rate of increase approaches zero at the highest levels of initial achievement.

In Figure 1 we also draw the 45 degree line that corresponds to the case where follow-up and baseline scores coincide. If $f(x) = x$ then the 45 degree line would be our best prediction of follow-up scores, and we would be back into the framework of a value-added specification that uses gain scores as the dependent variable.[8] Our estimated function is quite different than the 45 degree line and shows that we expect to see a sizable achievement increase for initially low-achieving students and little or no increase in scores at high levels of initial achievement. In fact, for the initial scores at the far right-tail of the baseline test score distribution, we actually expect that follow-up test scores will fall. Of course, this makes some sense since in the limiting case of a perfect initial score (60), our best predicted follow-up score should probably be something less than perfection. Finally, it is important to recognize that expected test score improvement is not constant across the initial achievement distribution, a result that may conflict with assumptions imposed in popular value-added specifications.

[7] Unless otherwise noted, we maintain the requirement that all schools in our sample must contain at least 5 students, though we will obtain results under a variety of sample selection rules.

[8] Of course, value-added models impose $f(x) = \delta_0 + x$, so that our estimate could depart from the 45 degree line by a constant and still be consistent with the value-added specification. However, the shape of our estimated regression function is clearly different than that used in value-added models.

4.2 Implications for Value-Added Models

The results of Figure 1 showed that our semiparametric estimate of the regression function relating intake and follow-up scores differed from the specification imposed by a value-added model. However, the fact that these relationships differ does not necessarily imply that key parameters of interest - such as the assessments of school performance and the subsequent determination of the roles for proxies for school quality - are affected by differences in the model specification. If assessments of school performance do vary across our models, this could potentially have significant implications regarding school performance assessment in developed countries - particularly the US - since the allocation of resources to schools can be based, in part, on rankings of schools by test score gains.[9]

In this section, we look at this issue in a bit more detail and examine how assessments of school performance in the Philippines are affected by changes in the models we employ. To do this we obtain estimated school effects for the value-added model: $y_{is} - x_{is} = \alpha_s^{va} + \epsilon_{is}^{va}$ and for the linear model: $y_{is} = \alpha_s^{lin} + \beta x_{is} + \epsilon_{is}^{lin}$ and compare these predictions to those obtained from our semiparametric hierarchical regression model in (3) and (4).

Ideally, we would compare predictions across models by coming up with a global measure to quantify the degree of similarity of our assessments of school performance. Though there are a variety of possible measures, we appeal to the *rank correlation* for this purpose.[10] Within each of our models we obtain a set of draws from the posterior distribution of the α_s parameters. For each iteration, we obtain a ranking of schools by sorting the $S \times 1$ (with S denoting the number of schools) vector of drawn α_s values. Doing this for every draw, we approximate the posterior distribution of the ranking of the school effects *within a given model*. Since there is uncertainty in the estimation of the α_s parameters, there will also be uncertainty in the ranking of schools - the best school will not necessarily have the highest α_s value at every iteration. However, the best schools in the sample will tend to be ranked highly across the iterations, while the worst schools will receive consistently low rankings.

Once the rankings are obtained *within each model*, we use the rank correlation to quantify the degree of similarity of our performance assessments *across our competing models*. For each iteration

[9] In California, for example, public schools can become eligible for the Governor's Reward Program by meeting an "'overall growth target." To meet this target, schools must improve their academic performance index (API) score by the larger of 5 points, or 5 percent of the difference between their base API and an overall target API. States are also given discretion in determining and rewarding school performance as a result of the "No Child Left Behind Act" of 2001. These designations are often made on the basis of test score gains.

[10] Use of the rank correlation is also relevant for the issue of funding schools on the basis of performance. Specifically, reward programs have been created with the intention of providing financial support to the best performing or highest ranked schools. As such, we look at the rank correlation to see how the ordering of schools varies across the value-added model and our semiparametric model. If the rankings differ, then our semiparametric model would predict a different allocation of resources to schools than is suggested by the often-used value added model.

(each draw from the posterior) we take the vector rankings from the semiparametric model and the corresponding vector of rankings for the value-added and linear models, and compute the rank correlation between them. The rank correlation (where there are no ties)[11] can be calculated as

$$r_s = 1 - \frac{6\sum_{i=1}^{S}(Se_i - Alt_i)^2}{S(S^2-1)},$$

where S denotes the total number of schools, Se_i denotes the semiparametric ranking of school i, and Alt_i denotes the ranking of school i under the alternate model (either value-added or linear). For each draw from the posterior distribution of the rankings, we obtain a numerical value for the rank correlation, r_s. Repeating this over all iterations enables us to obtain a posterior distribution for the rank correlation across our competing models.

Table 2
Posterior Means, Standard Deviations and Probabilities of Being Positive
for Rank Correlation with Semiparametric Model

	$n_s \geq 3$		$n_s \geq 5$		$n_s \geq 10$	
	Value-Added	Linear	Value-Added	Linear	Value-Added	Linear
Mean	.215	.445	.265	.525	.356	.654
Std.	.098	.087	.101	.083	.127	.078
$\Pr(\cdot > 0 \mid D)$	.980	1.00	.986	1.00	.994	1.00

Presented in Table 2 are posterior means, standard deviations and probabilities of being positive associated with the rank correlation r_s using our three different samples. As you can see, point estimates of the rank correlations are all positive, suggesting that schools ranked highly in our semiparametric model are also ranked highly in both the value-added and linear models. In addition, the posterior distribution of the rank correlation tends to put the vast majority of its mass over positive values, and no drawn value of the rank correlation between the semiparametric and linear model was ever negative. We see a much stronger correlation between the semiparametric and linear predictions than the semiparametric and value-added predictions. In fact, for some (albeit very few) of the iterations, the rank correlation between the value-added and semiparametric models was actually negative. Based on these results, we would expect to see a much stronger relationship between the linear and semiparametric predictions than between the value-added and semiparametric predictions. Intuitively, this makes sense as the semiparametric estimate in Figure 1 can be reasonably well approximated by a linear model, but clearly differs in shape from that imposed by the value-added specification.

The fact that the rank correlations in Table 2 are not close to one should not be interpreted as evidence that the semiparametric predictions are vastly different than either the value-added or linear

[11] In our context, there are no ties, as the α_s can always be ordered.

predictions. Even within a particular model, the correlation between rankings across iterations will not be close to unity simply because of uncertainty inherent in the estimation of the α_s parameters.

To get a different view of the similarity of predictions, we obtain a point estimate of our rankings for all three models and compare these point estimates across models.[12] The results of this exercise are presented in Figure 2. In this figure the semiparametric point estimates are plotted on the horizontal axis against the value-added (top) and linear (bottom) rankings on the vertical axis. Figure 2 presents these results for our largest sample with 93 schools, though similar results are obtained using our alternate sample restrictions. The rankings are ordered so that a ranking of 93 indicates the best performing school, while a ranking of 1 indicates the worst school. Finally, note that the case of perfect agreement in predictions corresponds to the case where the data points fall completely on the 45 degree line.

What we see from the figure is a tremendous agreement in rankings between the linear and semiparametric models, and a disparity of rankings between the semiparametric and value-added models. If we look a bit deeper at the source of the discrepancy between the value-added predictions and the semiparametric predictions, we find that the value-added models tend to strongly favor initially low achieving schools.

As suggestive evidence of this, consider for illustration points A and B in top portion of Figure 2. Point A corresponds to a school receiving a rank of (2,28) in the (semiparametric, value-added) models, while Point B corresponds to a school receiving a rank of (81,15) in the (semiparametric,value-added) models. Point A turns out to be an initially low achieving school with a mean intake score of 2.28 and a mean follow-up score of 10.12. Point B represents an initially high achieving school with a mean intake score of 27.06 and a mean follow-up score of 33.75. In terms of gain scores, school B improves less than school A, and so it tends to be ranked lower than school A under the value-added model that uses gain scores as the evaluation metric. However, our semiparametric model accounts for the fact that level improvements are not comparable across the initial achievement distribution, and recognizes that a 6.69 point average improvement is a significant accomplishment for an initially high achieving school. As a result, school B receives a high ranking of 81 under the semiparametric model. Conversely, the semiparametric model regards a 7.84 point improvement for an initially low achieving school as a relatively minor accomplishment (since most initially low achieving schools improve more than this), and assigns school A the second-worst ranking of 2.

Rather than looking at just two data points as suggestive evidence that value-added formulations strongly favor initially low achieving schools, we can use all the information provided in Figure 2 to document this result. To this end, we compute the difference between the value-added rankings

[12] In particular, we obtain the posterior mean.

and the semiparametric rankings for each school and regress these 93 differences on a constant and the *school-level mean intake score*. If the disparity in rankings between the value added and semiparametric models has nothing to do with school-level intake achievement, then we would expect an insignificant coefficient on the mean intake score that is close to zero. However, this regression produced a coefficient (and t-stat) for the intercept equal to 17.98 (-1.12), and for the slope equal to -1.12 (-14.27). The coefficient on the slope is highly significant, and suggests (roughly) that every point reduction in school-level mean intake scores tends to systematically increase the value-added ranking over the semiparametric ranking by 1. The R-squared for this regression was also .69, indicating that almost 70 percent of the variation in the gap between value-added and semiparametric rankings can be explained by variation in school-level mean intake scores.

These results suggest that the differences in regression function estimates provided in Figure 1 are important, as these differences generate very different assessments of school performance across the value-added and semiparametric models. In particular, value-added specifications tend to favor initially low-achieving schools. This finding has important implications not only for the evaluation of schools and for determining the roles of proxies for school quality in the Philippines, but also for school performance award programs in developed countries that are based largely on test score gains. From a statistical point of view, assessments of school performance from value-added specifications can not be rationalized as determining those schools that are performing better than expected. To properly model these expectations, one needs to recognize the possibility that level improvements are not comparable across the initial achievement distribution, and estimate a model that evaluates school performance on the basis of this estimated relationship.

4.3 Do Schools in the Philippines Matter?

The results in the previous sections have described one important feature of our model, as they presented flexible estimates of the relationship between baseline and follow-up test scores and documented important differences between our model and traditional value-added specifications. Another equally important goal of our model is to examine the magnitude and variation of the school effects α_s and to determine if schools in the Philippines really matter in the production of student achievement.

Table 3
Posterior Means and Standard Deviations for Variance Parameters

	$n_s \geq 3$		$n_s \geq 5$		$n_s \geq 10$	
	Post. Mean	Post Std.	Post. Mean	Post Std.	Post. Mean	Post Std.
σ^2_ϵ	128.82	(4.18)	129.17	(3.87)	130.86	(4.39)
σ^2_α	8.32	(2.60)	8.22	(2.62)	8.43	(2.84)
$\sigma^2_\alpha/(\sigma^2_\alpha + \sigma^2_\epsilon)$	.060	(.018)	.060	(.018)	.060	(.019)

Table 3 presents posterior means and standard deviations associated with the first and second-stage (respectively) variance parameters σ_ϵ^2 and σ_α^2. We present these point estimates under three different sample selection rules, which require at least 3,5, or 10 observations per school. In addition to simply presenting information about the variance parameters themselves, we also obtain the posterior distribution associated with the fraction of the total variation that is explainable at the school-level: $\sigma_\alpha^2/(\sigma_\alpha^2 + \sigma_\epsilon^2)$.[13] Across all of our samples, we find that 6 percent of the total conditional variation in follow-up scores is explainable by variation in performance across schools, and the posterior distribution of this fraction is rather tight and places little mass near zero. Specifically, no drawn value of $\sigma_\alpha^2/(\sigma_\alpha^2 + \sigma_\epsilon^2)$ was ever less than .025, suggesting that school-level variation accounts for at least 2.5 percent of the total variation in follow-up scores. This suggests some preliminary evidence that schools do indeed matter in determining student achievement and improvement, though it is important to recognize that the vast majority of conditional variation in performance arises at the individual level rather than the school level.

Our model and algorithm enables us to uncover much more than a simple characterization of the total variation across schools, and indeed, enables estimation of the posterior distribution of school-level effects (α_s) for every school in our sample.[14] In Figure 3 we present boxplots of all 72 of the school-level effects under the sample requiring at least 5 observations per school ($n_s \geq 5$). We have ordered the effects according to their posterior medians, and the interquartile ranges surrounding the medians have been plotted as wide boxes. The range of the posterior is given by the length of the solid vertical lines, and this segment connects the largest and smallest draws for each of the α_s effects.

As can be seen from the figure, the α_s distributions have an overall mean of zero, with the best schools having positive medians, and the worst schools having negative medians. In fact, if we use medians as our guide, we may expect to see as much as a ten point increase in test scores (roughly two-thirds of a standard deviation of the initial test score distribution) as a result of going from the worst to the best performing school.

4.3.1 Do Schools in the Philippines Really Matter? Some Additional Tests

There is a large amount of suggestive evidence from Figure 3 that schools in the Philippines play an important role in student achievement. The clear upward slope of the medians provides one bit of evidence supporting this claim, and the range between the best and worst schools is meaningfully

[13] The variance parameters σ_α^2 and σ_ϵ^2 we found to be somewhat negatively correlated. Our results simulate the exact finite sample distribution of this fraction by computing the value of $\sigma_\alpha^2/(\sigma_\alpha^2 + \sigma_\epsilon^2)$ for each draw.

[14] As shown in the appendix, the conditional posterior mean of each of the α_s parameters is essentially a weighted average of the OLS estimate using data from the given school and a common or pooled mean across all schools.

large. Further, many of the best schools have school-level effects that are positive with probability one (or approximately one), while the worst schools have school effects that are negative with probability one (or approximately one). Also note that the precision of the posteriors for the school effects (as represented by the length of the vertical lines) varies across schools, as the number of observations per school varies considerably in our data.

Though Figure 3 is suggestive of the importance of school-level effects, these results may not be fully satisfactory. To see why one may be skeptical, consider taking any data set, partitioning observations into groups and then estimating parameters for each group. Even if there are no true group effects, you will still see some variation in parameter estimates across groups. If we imagine generating data with a large error variance (as is the case in our analysis), and then partitioning the data into many groups, then seemingly one would estimate some groups to have large effects and others to have small effects even if there is no true group effect in the process generating the data. As such, we may not be certain if the results in Figure 3 reveal the empirical importance of schools, or if we would expect to see similar heterogeneity across groups even if there were no true group effect.

Though F-tests of joint equality provide an option for testing for the presence of school effects, their use is not ideal for our purposes. In our situation, we have a relatively large number of schools with some schools containing only a few observations, and it is typically easy in such situations to reject the null of joint equality in favor of the alternative where *at least two* of the schools are different. The finite-sample properties of such F-tests have been shown to be quite poor, and in particular, the use of F-tests in such situations has been shown to have a strong tendency to over-reject (*e.g* Bun (2003)).

Test #1: Reallocating Students to "Pseduo-Schools"

To address these concerns we consider two different and intuitive tests and apply them to our data. In the first test we keep the vector of school ID's as fixed and thus preserve the number of schools (72) as well as the distribution of observations per school that is present in our data. We then create "pseduo-schools" by randomly re-ordering our observations and assigning the reshuffled observations to the original school ID vector. This reallocation will not affect the shape of the function f, as the (x, y) pairs are preserved, nor will it affect the total error variance $(\sigma_\epsilon^2 + \sigma_\alpha^2)$. The fraction of the total variance that arises from across groups, however, certainly can change and this feature will provide a foundation for determining if school effects really are present in our data. If there is a true school effect, then we would expect to see larger estimates of σ_α^2 using the original structure of the data than is obtained when artificially reorganizing the data into groups of similar size.

To implement this test we randomly reshuffled the data with $n_s \geq 5$ (though similar results are obtained for the other samples) 100 times and fit the model with school effects for each of the 100

trials. We then recorded the posterior mean of σ_α^2 (which describes the extent of parameter variation at the group level) each time, and nonparametrically plot the distribution of the 100 obtained σ_α^2 point estimates. This distribution is presented in Figure 4. As you can see from Figure 4, none of the σ_α^2 point estimates ever exceeded .9, and no trial produced an estimate of σ_α^2 remotely close to the estimate we obtained using our original data, which was 8.22. This suggests that there are indeed important school effects, and no artificial reorganization of individuals to like groups produces comparable evidence of group effects.

Test #2: Null Simulations

We also consider and implement a second very similar test. In this test we take estimates of f, σ_ϵ^2 and σ_α^2 obtained from our original data and use these estimates to artificially generate y *under the null of no school effects*. Specifically, we generate $\tilde{y}_{is} = \hat{f}(x_{is}) + \hat{\epsilon}_{is}$, where $\hat{\epsilon}_{is} \overset{iid}{\sim} N(0, \hat{\sigma}_\alpha^2 + \hat{\sigma}_\epsilon^2)$ and $\hat{f}$ is obtained from our original data as in Figure 1. Once we have generated $\tilde{y}$, we take the $(x, \tilde{y})$ values and fit the model in (3)-(4) that incorporates school-level parameters into our estimation procedure. This experiment tells us the magnitude of variation across schools we would expect to see in a very similar design when there are no true school effects, *i.e.* the magnitude of variation across schools that arises purely from error in the regression function and the partitioning of our data into small groups.

As with the first experiment, we generate $\tilde{y}$ in this manner 100 times and record the posterior mean of σ_α^2 for each iteration. For the sake of brevity, we do not present our full set of results for this experiment, but note that the largest estimate of σ_α^2 in all 100 iterations was less than .25. This clearly indicates that the amount of school-level variation we find in our original data can not be explained as random error arising when one partitions the data into groups, and offers clear evidence against the null of no school effects. The estimation results of Table 3 together with the results of our two experiments lend strong and convincing support to the claim that schools are important determinants of student achievement in the Philippines.

4.4 Individual Controls and Measures of School Quality

In an effort to fix ideas our model has not yet added controls for individual characteristics other than the baseline test score (z_{is}), nor have we investigated the role of school quality variables (Q_s) in attempt to explain variation in school performance. In this section we examine the empirical importance of several additional controls at both the individual and school levels.

For the individual-level characteristics in z_{is} we add a dummy variable denoting if the student is

male (Male), number of years of schooling completed by the student's mother (MomEd) and father (DadEd) as of the base year (1994), and the log of household expenditure per household member (LogExpend). For our set of school quality variables, we choose rather parsimonious specifications due to the limited number of schools in our data. The variables we include, however, seem to be of primary importance and overlap with those employed in other studies.

Specifically, we include ClassSize, defined as the total number of students in the school divided by the number of sessions taught, the percentage of teachers in the school who received in-service training (InService), the average experience of teachers in the school (TeachExp) and a dummy denoting if the school has electricity (Electricity).[15] All of the quality variables are standardized to have mean zero and unit variances for purposes of identification[16] and interpretation.

To investigate the sensitivity of our results we obtain point estimates of these coefficients under a variety of sample selection rules. In particular, we obtain results under our maintained restriction that there must be at least 5 students per school, and also obtain estimates when we require at least 3 or 10 students per school. In all of these samples, the average number of students per school is reasonably large, and equals 21.9 in the largest sample ($n_s \geq 3$) and 38.3 in the smallest sample ($n_s \geq 10$). As shown in Table 4, point estimates and variable significance were largely unaffected by the changes in the samples.

[15] We repeated our analysis to include the facilities index variable, FacIndx, and found this index to be a significant predictor of test scores. However, the index appeared to be driven primarily by the availability of electricity, and so we simply include the electricity variable in our reported specifications. The variable BookStu (no of books in the library per student) was not significant in most of the specifications with controls for Electricity.

[16] This is done so that the common mean of the school-level effects is zero to avoid confounding an intercept associated with the school effects with an intercept embedded in the function f.

Table 4
Coefficient Posterior Means, Standard Deviations and
Probabilities of Being Positive Under Alternate Sample Restrictions

	Sample Restriction $n_s \geq 3$			Sample Restriction $n_s \geq 5$			Sample Restriction $n_s \geq 10$		
	Post Mean	Post Mean	$\Pr(\cdot > 0\|D)$	Post Mean	Post Mean	$\Pr(\cdot > 0\|D)$	Post Mean	Post Mean	$\Pr(\cdot > 0\|D)$
X Variables									
Male	-3.22	(.476)	.000	-3.24	(.510)	.000	-3.41	(.534)	.000
MomEd	.120	(.081)	.945	.117	(.078)	.935	.102	(.092)	.875
DadEd	.167	(.069)	.945	.177	(.067)	.995	.202	(.076)	.95
LogExpend	.765	(.265)	1.00	.829	(.179)	1.00	.988	(.283)	1.00
σ^2_ϵ	123.9	(4.06)	1.00	124.3	(3.96)	1.00	125.7	(4.22)	1.00
Q Variables									
ClassSize	.468	(.485)	.805	.119	(.490)	.595	-.005	(.548)	.501
InService	.057	(.540)	.528	.102	(.456)	.587	.076	(.554)	.565
TeachExp	-2.02	(.483)	.000	-2.18	(.502)	.000	-2.11	(.587)	.000
Electricity	1.54	(.527)	.995	1.91	(.547)	1.00	2.01	(.642)	1.00
σ^2_α	5.30	(2.08)	1.00	3.89	(1.72)	1.00	5.22	(1.92)	1.00
$\frac{\sigma^2_\alpha}{\sigma^2_\alpha+\sigma^2_\epsilon}$	.041	(.016)	1.00	.030	.013	1.00	.040	(.014)	1.00
n	2,038			1,970			1,801		
S	93			72			47		

Based on the results presented in Table 4, we find convincing evidence that family characteristics are influential in determining achievement growth even after flexibly controlling for initial test scores. Specifically, parental education and log per-capita household expenditure were found to be important determinants of follow-up scores in all of our results, and the coefficients associated with these variables were found to be positive with probabilities equal to one or close to one. An additional year of parental education increased follow-up scores from .1-.2 points across our various samples. Evaluated at mean values, a 1,000 Philippine-peso increase in household expenditure per family member was associated with a test score increase ranging from .24 - .33 points. Finally, we also note that the Male dummy variable is strongly significant across our samples, indicating that, on average, female students in the Philippines improved by about 3.2-3.4 points more than male students.

The empirical importance of parental education and household resources suggests some potential for error in school performance assessment even when analyses are performed that condition on intake achievement. The implicit assumptions made in these analyses is that school performance can be elicited empirically by looking at changes in test scores or upon successfully controlling for nteresting results. First, the electricity coefficient is strongly positive across all three of our samples and suggests that schools with electricity perform much better than schools without it. Specifically, schools with electricity tend to outperform other schools by about 4.28-4.90 points across our various samples. [17]

[17]Recall that the quality variables are standardized so care must be taken in interpreting the meaning of the

These are large differences and suggest strong evidence that the availability of basic school resources plays an important role in explaining variation in school performance in the Philippines. While one could argue that this finding is just picking up a selection effect, as the best students in the Philippines are also the ones that attend schools with electricity, it is important to note that this result is obtained after controlling for parental education and household expenditure and evaluating schools on the basis of *improvements* in test scores rather than level achievements. If we look a bit more deeply at our data with $n_s \geq 5$ we find that *only one* of the top 34 schools (as ranked by the posterior medians of the α_s parameters) do not have electricity, and all of the schools in the top 10 had electricity. At the other end of the rankings distribution, we found that only one of the worst 5 schools reported to have electricity. For the largest sample with $n_s \geq 3$, 43 of the top 44 schools and only 1 of the bottom 5 schools had electricity.

In addition to the provision of electricity we find a strong association between average age of teachers in the school and school performance. Specifically, a one-standard deviation increase in the average age of teachers in the school (which ranges from an increase of 6.5 to 7 years across our three samples) leads to reduction in test scores ranging from 2 to 2.2 points. For our largest sample with $n_s \geq 3$ we find that only 5 of the top 30 schools had average teacher experience levels greater than the sample mean and only 7 of the bottom 30 schools had average teacher experience levels less than the sample mean. As mentioned in our discussion of the data, teachers in our sample tend to be older and possess approximately 14-15 years of experience on average. Given this, we might interpret our findings as a local result that reveals a negative effect of additional teacher experience for a sample of teachers who are already reasonably experienced.

While these results clearly suggest the importance of basic school resources and teacher experience, more "traditional" school quality measures like ClassSize and teacher training play virtually no role in explaining variation in performance at the school level. This finding is consistent with several other studies of schooling in developing countries, such as Harbison and Hanushek (1992), that find strong evidence that facilities of the school are important determinants of student achievement but characteristics of the teachers and student-teacher ratios play relatively little role in explaining variation in school performance.

It is also important to recognize that our model has uncovered these results after controlling for nonlinearity in the relationship between baseline and follow-up scores. As discussed above, the estimated school-level parameters α_s we obtain in our model can be quite different from those obtained under alternate models that impose particular structure on the relationship between baseline and follow-up scores. As such, we could reach very different conclusions regarding the empirical impor-

coefficients. The standardized electricity dummy variable is still binary, but is not $\in \{0, 1\}$ due to the standardization. For the case of continuous covariates, marginal effects are calculated and reported using the chain rule: $\partial y/\partial Q = (\partial y/\partial \tilde{Q})(\partial \tilde{Q}/\partial Q)$, where $\tilde{Q}$ represents the standardized quality variable.

tance of proxies for school quality, since they are added to explain variation in these school-level effects. Our results, however, seem to echo those of previous studies. In particular we find that the provision of basic school resources (particularly electricity) is important in student achievement growth, but other variables such as class size and teacher training play little role in explaining variation in performance across schools.

5 Conclusion

An extensive empirical literature that studies the relative importance of schooling and family inputs on children's test score outcomes makes use of the education production function. When test score data are available at two points in time, the specifications typically estimated include the value-added approach - models that use the change in test scores as the dependent variable - or models that regress current test score on lagged test scores. Using a rich set of data from Cebu, Philippines, we introduce and estimate a semiparametric hierarchical model of achievement growth that nests both of these models, and find that the data rejects a linear relationship between the baseline and follow-up test score. Accurately describing this relationship is especially crucial in identifying if there is systematic variation in the effect of schools on student achievement.

Our study contributes not only to the general issue of accurately modeling and estimating the education production function–which could be applied to both developed and developing country settings–but also provides evidence of the impact of schools on the achievement of students from a developing country. Most previous work in developing countries are typically faced with data at one point in time only, and regress educational outcomes on a host of contemporaneous school and family inputs while controlling for school choice. Our approach differs from this in that we have the requisite data to control for the effect of historical inputs, and we generally identify the effect of schools in the production of achievement growth.

Given compelling evidence that schools differ, we looked for measurable school characteristics that were capable of explaining school-level performance variation. Our results are consistent with an emerging theme among recent studies in developing countries. Minimal basic facilities, and particularly in Cebu, the provision of electricity, mattered more than class size, or teacher training programs. Unlike models employed in previous studies, our model obtains estimates of the relative importance of school quality variables after controlling for potential nonlinearity in the relationship between baseline test scores and follow-up test scores. Successfully accounting for this nonlinearity could have resulted in very different predictions regarding the empirical importance of the school quality variables.

On the other hand, although we do find that schools clearly matter, only 6 percent of the total variation in follow-up scores can be explained by variation in performance across schools. Thus, most of the variation in follow-up achievement stems from individual-level characteristics that are unobserved. This suggests that policies implemented in developing countries to stimulate improvements in human capital should not only be targeted at schools but also at households.

6 Appendix A: Figures

Figure 1: Nonparametric Estimate of Expected Follow-up Score given Baseline Score

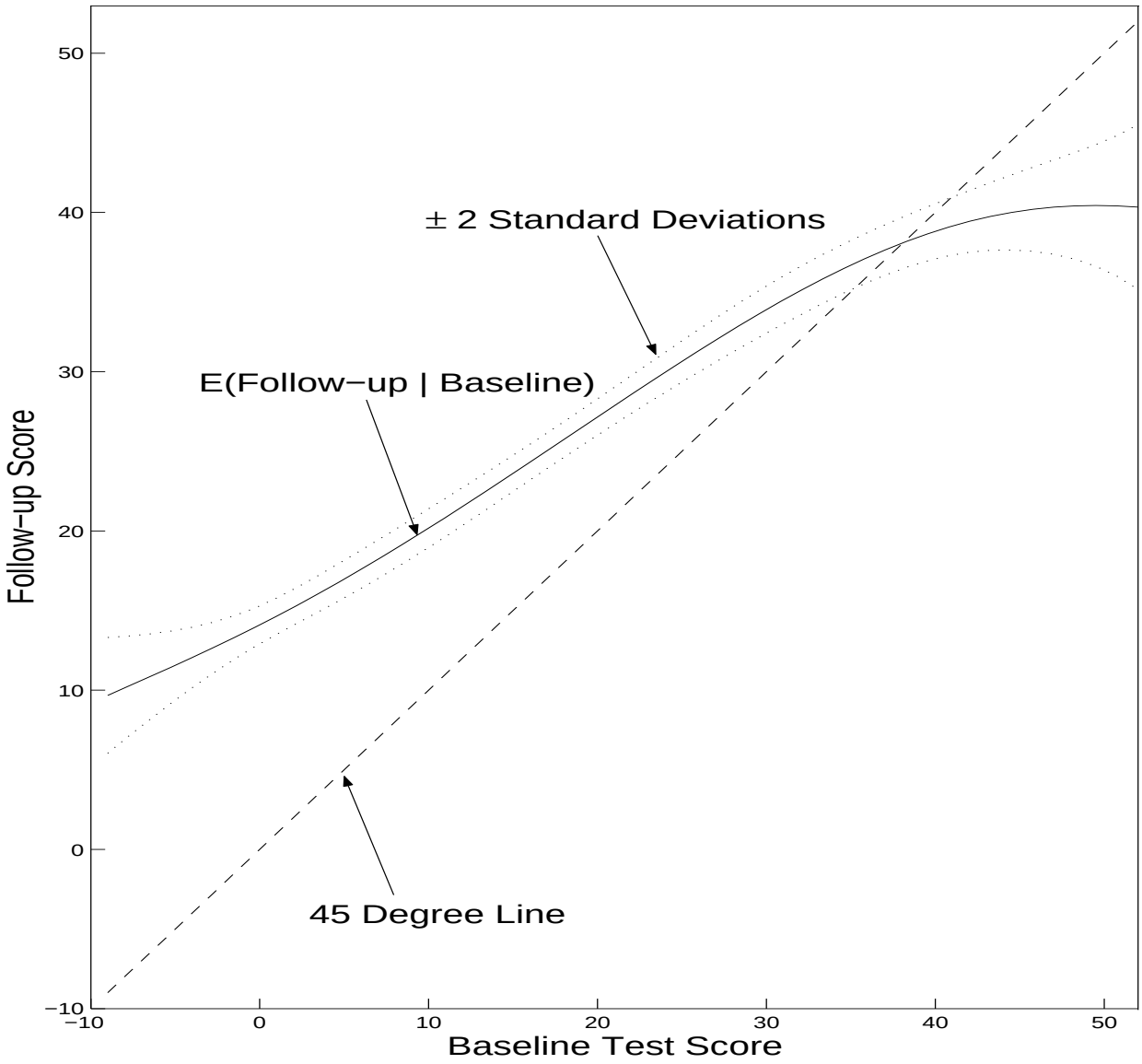
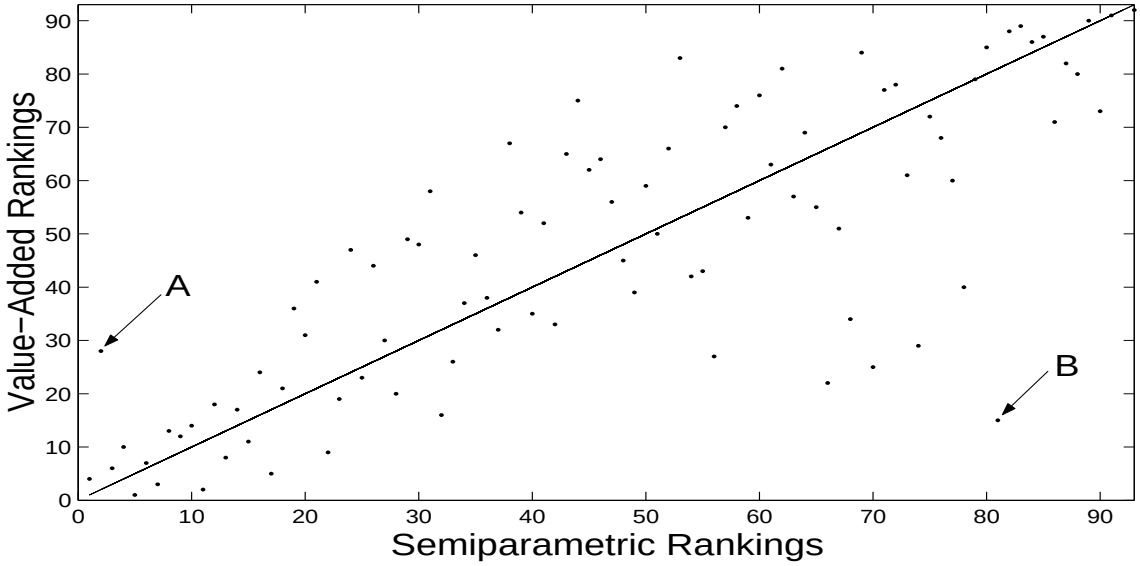


Figure 2: Comparison of Point Estimates of School Rankings Across Various Models



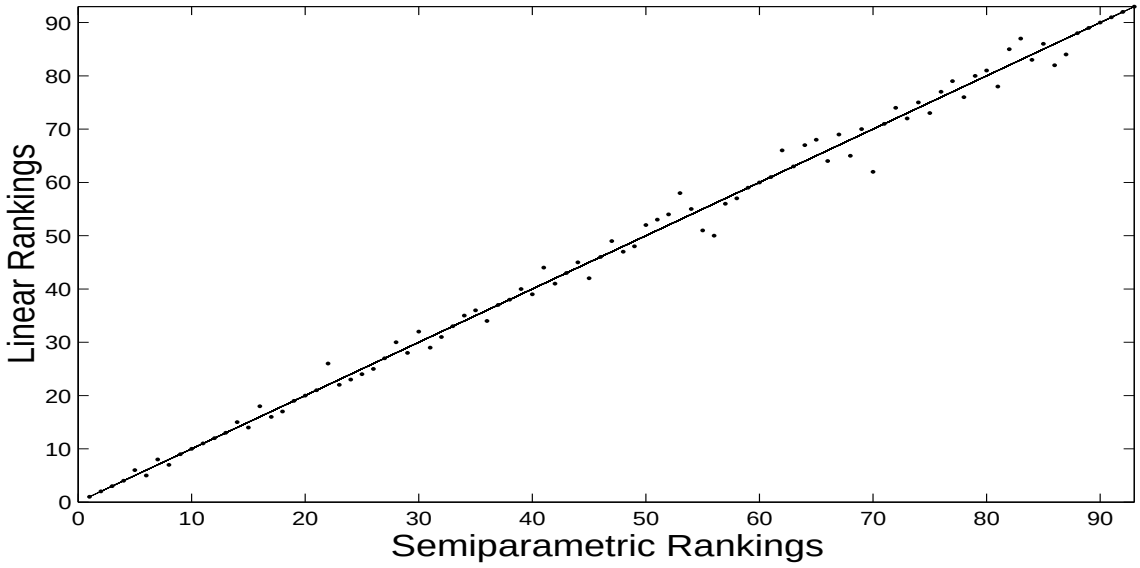


Figure 3: Boxplot of 72 School-Specific Random Effect Posteriors ($p(\alpha_s|\text{Data})$), Ordered by their Posterior Medians

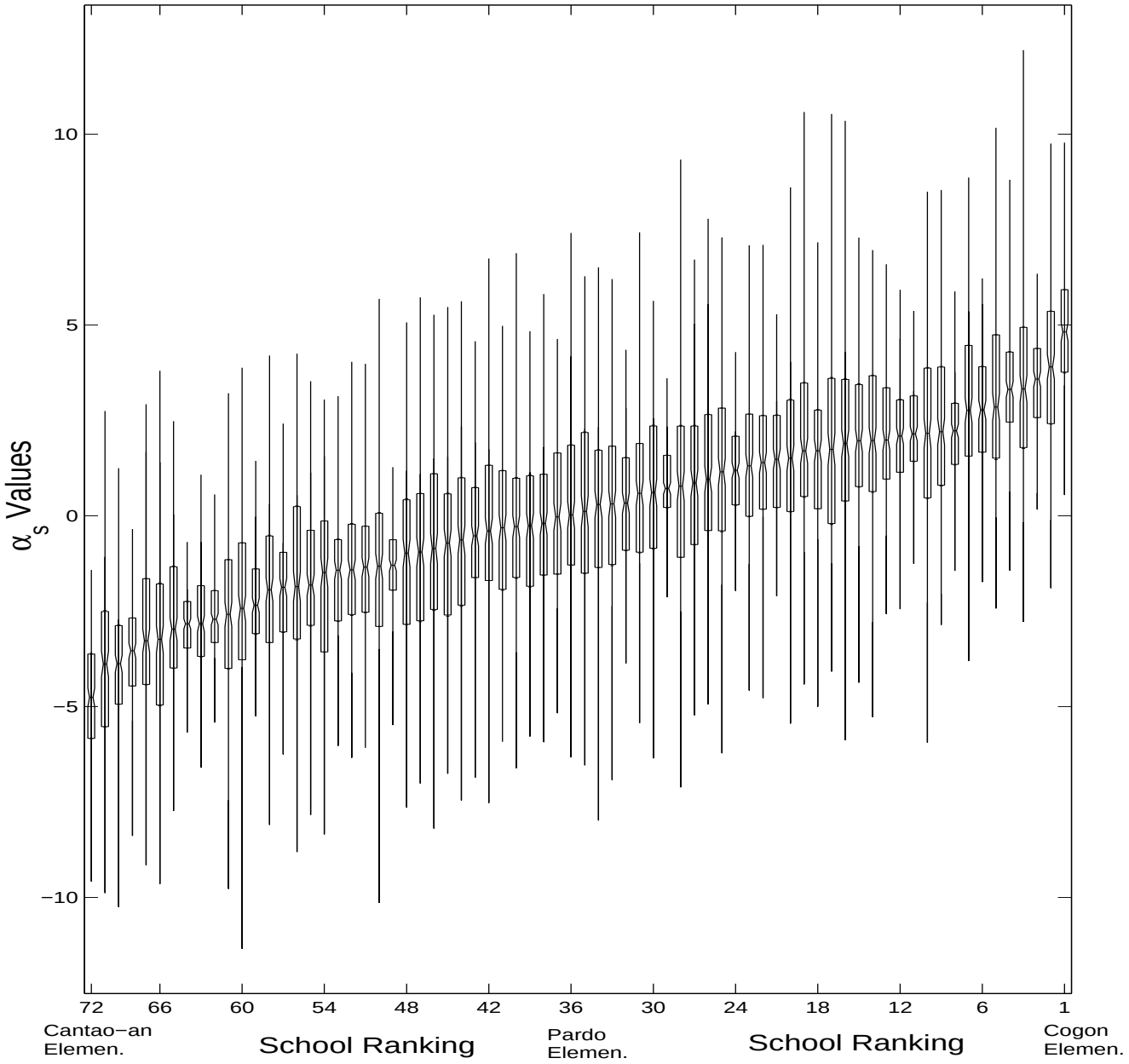
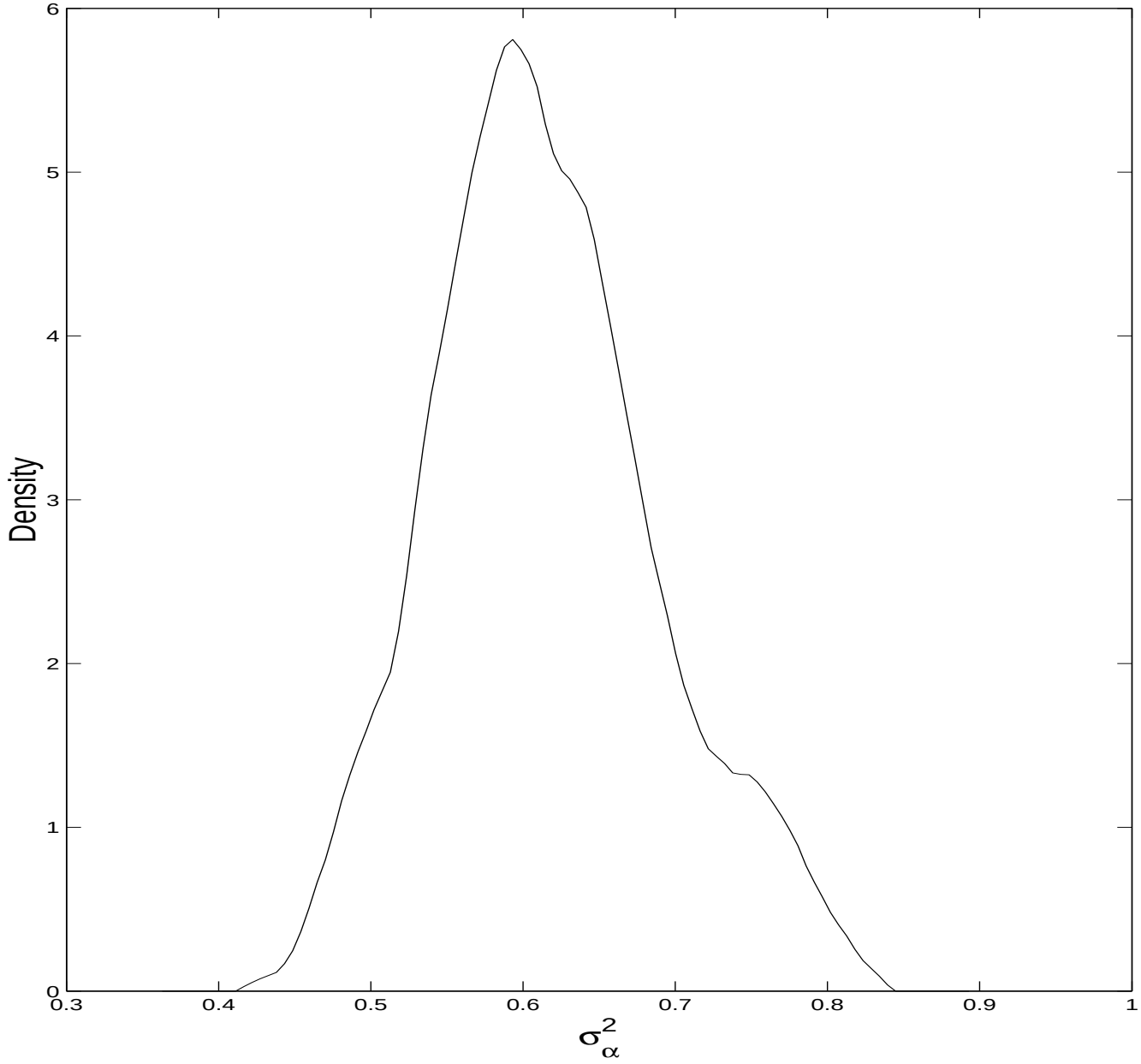


Figure 4: Density of $\hat{\sigma}^2_\alpha$ values obtained by Randomly Reallocating Individuals to "Pseudo=Schools"



7 Appendix B: The Posterior Simulator

Our most general model, as presented in (1) - (2), is given by:

$$\begin{aligned} y_{is} &= \alpha_s + f(x_{is}) + z_{is}\beta + \epsilon_{is}, \qquad \epsilon_{is} \overset{iid}{\sim} N(0, \sigma_\epsilon^2) \\ \alpha_s &= Q_s\gamma + u_s, \qquad\qquad\qquad\quad u_s \overset{iid}{\sim} N(0, \sigma_\alpha^2). \end{aligned}$$

We employ a Bayesian estimation approach and fit the model above using the *Gibbs sampler* - an iterative simulation scheme that successively samples from the complete posterior conditional distributions. Our method for nonparametric estimation of f uses a cubic regression spline:

$$f(x) = \delta_0 + \delta_1 x + \delta_2 x^2 + \sum_{j=3}^{J} \delta_j (x - \tau_{j-2})_+^3,$$

where $z_+ \equiv \max(0, z)$, and $\{\tau_j\}_{j=1}^{J-2}$ denote our potential knot points placed in the interior of the support of x with $\tau_j < \tau_{j+1}$.

As in Smith and Kohn (1996) we regard the problem of knot point determination as a problem of variable selection. To this end we let $X_{n\times(J+1)} = [\mathbf{1}\ x\ x^2\ (x-\tau_1)_+^3\ \cdots\ (x-\tau_{J-2})_+^3]$ and will seek to determine those columns of X that are needed to accurately estimate the regression function f. We let $\theta = [\theta_1\ \theta_2\ \cdots\ \theta_{J+1}]$ be an *indicator vector* denoting the columns of X that are to be included in the model, define X_θ as the restricted X matrix that contains only those columns of X that are kept in the model given θ, and define δ_θ as the regression coefficient vector associated with X_θ. The method we outline below analytically determines the posterior probability associated with each possible combination of retaining/excluding the various columns of X (2^{J+1} possible combinations), and thereby finds the spline function within this flexible class that provides the best description of the regression function f. In practice we allow for 15 potential knot points and equally space them throughout the baseline test score support $[-8, 53]$. Results were not affected by allowing for a few additional knot points or by rearranging the location of the potential knots.

To complete our Bayesian analysis, we specify priors for our model parameters of the following forms:

$$\begin{aligned} \delta_\theta | c, \theta, \sigma_\epsilon^2 &\sim N\left(0, c\sigma_\epsilon^2 (X_\theta' X_\theta)^{-1}\right) \\ p(\sigma_\epsilon^2) &\propto 1/\sigma_\epsilon^2 \\ \theta | \pi &= \prod_{j=1}^{J} \pi_j^{\theta_j} (1-\pi_j)^{1-\theta_j} \\ \beta | \mu_\beta, V_\beta &\sim N(\mu_\beta, V_\beta) \\ \gamma | \mu_\gamma, V_\gamma &\sim N(\mu_\gamma, V_\gamma) \\ \sigma_\alpha^2 | a, b &\sim IG(a, b). \end{aligned}$$

In practice, we set $c = n$ so that our prior for the coefficient vector δ_θ is quite diffuse and approximates the information that would be contained in only one observation. We choose a non-informative improper prior for σ_ϵ^2, and set $\pi_j = 1/2 \; \forall j$ so all potential explanatory variables are equally likely to be included *a priori*. Finally, we set $\mu_\beta = \mu_\gamma = 0$, and choose V_β and V_γ as diagonal matrices with diagonal elements equal to 100, so that the prior standard deviation of each coefficient is large and the data information will be predominant. For the parameters of our Inverted Gamma (IG) density for σ_α^2, we set $a = 3$ and $b = 1/2$, which sets the prior mean of σ_α^2 equal to 1, and also specifies a relatively diffuse specification for the school-level variance around this mean. As shown in section 4, the posterior pulls us away from this rather conservative prior, and reveals substantially more school-level heterogeneity than is suggested in our prior.

8 The Blocked Gibbs Algorithm

The parameters of our model consist of $\Gamma = [\theta \; \delta_\theta \; \sigma_\epsilon^2 \; \{\alpha_s\} \; \gamma \; \sigma_\alpha^2]$. We use the Gibbs sampler to fit this model, and thus need to determine the complete posterior conditional distributions. We offer an improvement over the standard Gibbs sampler and use several *blocking steps* to mitigate the effect of autocorrelation in our parameter chains. In particular, we cycle though the following blocked posterior conditionals:

$$\begin{aligned}
&p(\theta, \delta_\theta, \sigma_\epsilon^2 | \Gamma_{-\theta,\delta_\theta,\sigma_\epsilon^2}, \text{Data}) \\
&p(\{\alpha_s\}, \beta | \Gamma_{-\{\alpha_s\},\beta}, \text{Data}) \\
&p(\gamma | \Gamma_{-\gamma}, \text{Data}) \\
&p(\sigma_\alpha^2 | \Gamma_{-\sigma_\alpha^2}, \text{Data}),
\end{aligned}$$

with Γ_{-x} denoting all parameters other than x.

First Block: $p(\theta, \delta_\theta, \sigma_\epsilon^2 | \Gamma_{-\theta,\delta_\theta,\sigma_\epsilon^2}, \text{Data})$

We draw from the joint conditional posterior of $\theta, \delta_\theta, \sigma_\epsilon^2$ by drawing from the marginal conditional for θ, plugging this draw in to the conditional for σ_ϵ^2 given θ and the parameters other than δ_θ, and then drawing from the complete conditional for δ_θ. With a bit of work, one can obtain the following expressions for these posterior conditionals:

$$\theta | \Gamma_{-\theta,\delta_\theta,\sigma_\epsilon^2}, \text{Data} \propto (1+c)^{(-q_\theta/2)} \left[\tilde{y}'\tilde{y} - (c/(1+c)) \tilde{y}' X_\theta' (X_\theta' X_\theta)^{-1} X_\theta' \tilde{y} \right]^{-n/2},$$

where $\tilde{y} \equiv y - \overline{\alpha} - Z\beta$.

$$\sigma_\epsilon^2 | \Gamma_{-\delta_\theta,\sigma_\epsilon}, \text{Data} \sim IG\left(\frac{n}{2}, 2\tilde{y}(I_n - (c/(1+c)) X_\theta (X_\theta' X_\theta)^{-1} X_\theta') \tilde{y} \right).$$

$$\delta_\theta|\Gamma_{-\delta_\theta}, \text{Data} \sim N\big[(c/(1+c))(X_\theta'X_\theta)^{-1}X_\theta'\tilde{y}, \sigma_\epsilon^2(c/(1+c))(X_\theta'X_\theta)^{-1}\big].$$

In the above we let $\overline{\alpha}$ denote the $n \times 1$ vector of school effects blocked by school and define q_θ as the number of columns in X_θ. We also note that θ is a discrete-valued vector of zeros and ones, and thus we can evaluate the posterior $p(\theta|\Gamma_{-\theta,\delta_\theta,\sigma_\epsilon^2}, \text{Data})$ for all possible values of θ and then can draw from this discrete distribution.

Second Block: $p(\{\alpha_s\}, \beta|\Gamma_{-\{\alpha_s\},\beta}, \text{Data})$

To draw from the second blocked group, we draw β from its conditional posterior marginalized over the school effects, and then draw the $\{\alpha_s\}$ successively, conditioned on the value of β from the previous step. Specifically:

$$\beta|\Gamma_{-\beta,\{\alpha_s\}}, \text{Data} \sim N(D_\beta d_\beta D_\beta),$$

where

$$\begin{aligned} D_\beta &= \left(\sum_{s=1}^{S} Z_s'V_s^{-1}Z_s + V_\beta^{-1}\right)^{-1}, \quad d_\beta = \left[\sum_{s=1}^{S} Z_s'V_s^{-1}(y_s - i_{n_s}Q_s\gamma - X_\theta\delta_\theta)\right] + V_\beta^{-1}\mu_\beta \\ V_s^{-1} &\equiv \sigma_\epsilon^{-2}\left[I_{n_s} - \frac{\sigma_\alpha^2}{\sigma_\epsilon^2 + n_s\sigma_\alpha^2} i_{n_s}i_{n_s}'\right], \end{aligned}$$

where S denotes the total number of schools (and Z_s, V_s, Q_s and y_s are defined accordingly), n_s denotes the number of observations in school s, and i_k denotes a $k \times 1$ vector of ones. Then,

$$\alpha_s|\Gamma_{-\alpha_s}, \text{Data} \overset{ind}{\sim} N(D_{\alpha_s}d_{\alpha_s}, D_{\alpha_s}),$$

where

$$D_{\alpha_s} = \left[n_s/\sigma_\epsilon^2 + 1/\sigma_\alpha^2\right]^{-1}, \quad d_{\alpha_s} = \left[\sum_{i\in s}(y_{is} - x_{\theta_{is}}\delta_\theta - z_{is}\beta)/\sigma_\epsilon^2\right] + (1/\sigma_\alpha^2)Q_s\gamma.$$

The remaining two complete posterior conditionals are obtained:

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

where

$$D_\gamma = \left(Q'Q/\sigma_\alpha^2 + V_\gamma^{-1}\right)^{-1}, \quad d_\gamma = Q'\alpha/\sigma_\alpha^2 + V_\gamma^{-1}\mu_\gamma.$$

and

$$\sigma_\alpha^2|\Gamma_{-\sigma_\alpha^2}, \text{Data} \sim IG\left(S/2 + a, \left[b^{-1} + .5(\alpha - Q\gamma)'(\alpha - Q\gamma)\right]^{-1}\right),$$

where $\alpha \equiv [\alpha_1 \ \alpha_2 \ \cdots \ \alpha_S]'$. The Gibbs sampler is run for 10,000 iterations and the final 9,000 iterations are used to calculate posterior means standard deviations, and other quantities of interest. Our experiments showed that the blocking steps significantly help to facilitate convergence, and parameter chains appeared to settle down within 100 iterations.

9 Appendix C: Sample Selection and Attrition from the CLHNS

Live Births in 33 Sample Barangays of Metro Cebu		3,289	
Of which:	Twin Births	27	(0.8%)
	Refusals	97	(2.9%)
	Missed by Survey (discovered later)	58	(1.8%)
	Birth Interview Too Late	22	(0.7%)
Live Births in Metro Cebu with Birth Interview		3,085	
Of which:	Migrated Out of Metro Cebu by Age 2	318	(10.3%)
	Child Died by Age 2	156	(5.1%)
	Refusal (at later date)	50	(1.6%)
Still in Sample When Child is 2 years old		2,561	
Of which:	Migrated Out of Metro Cebu by Age 8	155	(6.1%)
	Could not find child at Age 8	137	(5.3%)
	Child Died by Age 8	38	(1.5%)
Still in Sample When Child is 8 yrs old (1991-92)		2,231	
Of which:	Migrated Out/Could Not Find	31	(1.4%)
	Child Died	8	(0.4%)
Still in Sample When Child is 11 yrs old (1994-95)		2,192	
Of which:	Never Enrolled in School	9	(0.4%)
	Not Tested (refusal)	13	(0.6%)
	Had Younger Sibling of School Age	1,261	(57.5%)
Total Observations in 1996-97	Index & Sibling Children with Test Scores	3,258	
Of which:	Not enrolled 1996/97	282	(8.66%)
	Transferred schools bet 1994/95 & 1996/97 (exceeded max grade offered at previous school, e.g., moved from elementary to a high school)	616	(18.9%)
	Transferred schools bet 1994/95 & 1996/97 (other reasons)	224	(6.9%)
Observations in analysis		2,136	

References

[1] Behrman, Jere R. and Nancy Birdsall. 1983. "The Quality of Schooling: Quantity Alone is Misleading," *American Economic Review* 73(5): 928-946.

[2] Betts, Julian. 1995. "Does School Quality Matter? Evidence from the NLSY," *The Review of Economics and Statistics*, 77 (May): 231-250.

[3] Bun, M.J.G. (2003). "Testing Poolability in a System of Dynamic Regressions with Nonspherical Disturbances. mimeo.

[4] Card, David and Alan Krueger. 1992. "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States," *Journal of Political Economy*, 100(February): 1-40.

[5] Chib, S. and I. Jeliazkov. 2003. Semiparametric Hierarchical Bayes Analysis of Discrete Panel Data with State Dependence. Under review, *Econometrica*.

[6] Coleman, James, E. Campbell, et.al. 1966. Equality for Educational Opportunity U.S. Department of Health, Education, and Welfare, Office of Education. Washington, D.C.: U.S. Government Printing Office.

[7] DiNardo, J. and J.L. Tobias 2001. Nonparametric Density and Regression Estimation. *Journal of Economic Perspectives* 15(4), 11-28.

[8] Duflo, Esther. 2001. "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment," *American Economic Review*, 91(4): 795-813.

[9] Ehrenberg, R. and D. Brewer. 1994. "Do School and Teacher Characteristics Matter? " *Economics of Education Review* 13: 291-310.

[10] Fan, J. and I. Gijbels. 1996. *Local Polynomial Modeling and its Applications* London: Chapman and Hall.

[11] Fuller, Bruce and Prema Clark. 1994. "What School Factors Raise Achievement in the Third World?" *Review of Education Research*, 57(3): 255-92.

[12] Green, P.J. and B.W. Silverman. 1994. *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach.* London: Chapman and Hall.

[13] Glewwe, Paul. 2002. "Schools and Skills in Developing Countries: Education Policies and Socioeconomic Outcomes," *Journal of Economic Literature*, 40 (June), pp 436-82.

[14] Grogger, Jeff. 1996. "School Expenditures and Post-Schooling Earnings: Evidence from the High School and Beyond." *The Review of Economics and Statistics*, 78(4): 628-37.

[15] Hanushek, Eric. 1986. "The Economics of Schooling: Production and Efficiency in Public Schools." *Journal of Economic Literature*, 24(3): 1141-1177.

[16] Hanushek, Eric. 1995. "Interpreting Recent Research on Schooling in Developing Countries," *World Bank Research Observer*, 10(2): 227-46.

[17] Hanushek, Eric. 2002. "The Failure of Input-Based Schooling Policies," *Economic Journal*, 113 (485): F64-F98.

[18] Hanushek, Eric, John Kain, Steve Rivkin. 1998. "Teachers, Schools, and Academic Achievement," NBER Working Paper No. 6691, August.

[19] Hanushek, Eric, John Kain, Jacob Markman, Steve Rivkin. 2001. "Does Peer Ability Affect Student Achievement?," NBER Working Paper No. 8502, October.

[20] Harbison, Ralph W. and Eric A. Hanushek. 1992. Educational performance of the poor: Lessons from rural northeast Brazil. New York: Oxford University Press for the World Bank.

[21] Kane, Thomas and Douglas Staiger. 2002. "The Promise and Pitfalls of Using Imprecise School Accountability Measures," *Journal of Economic Perspectives*, Fall, pp 91-114.

[22] Kingdon, Geeta. 1996. "Student Achievement and Teacher Pay: A Case Study of India," STICERD working paper 74, London School of Econ. Political Science.

[23] Koop, G. and D. Poirier 2002. Bayesian Variants of Some Classical Semiparametric Regression Techniques. Under review, *Journal of Econometrics*.

[24] Kremer, Michael. 1995. "Research on Schooling: What we know and what we don't, A Comment on Hanushek," *World Bank Research Observer*, 10(2): 247-54.

[25] 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.

[26] Office of Population Studies. 1989. "The Cebu Longitudinal Health and Nutrition Study: Survey Procedures and Instruments". University of San Carlos. Cebu City. Philippines.

[27] Smith, M. and R. Kohn. 1996. Nonparametric Regression Using Bayesian Variable Selection. *Journal of Econometrics* 75, 317-343.