Schools, School Quality and Achievement Growth: Evidence From the Philippines[1]

June 29, 2005

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 explains about 4.4 to 5.3 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]We would like to thank Jun Ishii, Dale Poirier, Steve Rivkin, Sarah Senesky and two anonymous referees for helpful comments and discussion related to this paper. All errors are, of course, our own.

1 Introduction

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 primarily adopted one of two estimation approaches. The "value-added" approach assesses school performance by modeling changes in test scores (i.e., 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 models assuming linearity in the relationship between intake and follow-up scores. Formally, the model we estimate is a *semiparametric hierarchical regression model.*

For each student in our data with a given intake test 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 should appeal to other researchers seeking to evaluate school performance. To see the potential usefulness of such an approach, it is important to first 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.[2] 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 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).

We find compelling evidence that schools in the Philippines differ 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

[2] 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.

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-1990's, 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 determinants of student achievement in the Philippines, we find that schools account for only 4.5-5.3 percent of the (conditional) 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:

$$
\begin{aligned}
y_{is} &= \alpha_s + f(x_{is}) + z_{is}\beta + \epsilon_{is}, \qquad \epsilon_{is}|x_{is}, z_{is} \overset{iid}{\sim} N(0, \sigma_\epsilon^2) \qquad (1) \\
\alpha_s &= Q_s\gamma + u_s, \qquad u_s|Q_s \overset{iid}{\sim} N(0, \sigma_\alpha^2). \qquad (2)
\end{aligned}
$$

In the above, y_{is} refers to the follow-up test score 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.

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 $f(x) = \delta_0 + \delta_1 x$.

School effects are captured in the α_s parameters which 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 are 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 regression model in (1) and (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 this semiparametric hierarchical specification.

Estimation of the Nonparametric Component

To be sure, there are a number of possible 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), Koop and

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

Tobias (2005), 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 quantity as a parameter to be determined within our model. The algorithm is *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. MATLAB code and the data for fitting this model are available upon request, and specific details regarding our algorithm and the priors employed can be found in Appendix B.

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

[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.

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.

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

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

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 dummy 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 cancelled 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 equations (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

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

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, we estimate a simplified version of our model in (1) and (2):[7]

$$\begin{aligned} y_{is} &= \alpha_s + f(x_{is}) + \epsilon_{is}, \qquad \epsilon_{is} \overset{iid}{\sim} N(0, \sigma^2_\epsilon) \qquad (3) \\ \alpha_s &\overset{iid}{\sim} N(0, \sigma^2_\alpha). \qquad (4) \end{aligned}$$

We begin with this restricted model since we first want to explore the relationship between intake test scores and follow-up test scores and examine if nonlinearities are present in this relationship. This restricted model also serves as a useful starting point since in many situations, schools are evaluated solely on the basis of test score improvements over time and no additional student or school level controls are accounted for when obtaining estimates of school performance.[8]

4.1 The Shape of the Regression Function

In Figure 1 we present our point estimates and standard error bands associated with the function $f(x)$ in (3) which represents the conditional expectation of follow-up test scores given the initial (intake) test score. The estimated function is clearly increasing throughout the support of initial test scores, indicating that higher baseline scores lead to higher expected follow-up scores. In addition, it is not possible to draw a straight line within the standard error bands, providing evidence against a purely linear specification.

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.[9] Our estimated function is quite different than the 45 degree line and shows a sizable expected achievement increase for initially low-achieving students and little or no increase in scores at high levels of initial achievement.[10]

[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]See, for example, Tobias (2004).

[9]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.

[10]We find that the regression function f has a similar shape even when student-level controls such as parental

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 the value-added and linear models. However, the fact that these relationships differ does not necessarily imply that key quantities 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.

In this section, we look at this issue in 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. 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.

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 (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).

education and family resources are included in equation (1). Thus, the substantive points discussed in this section generalize to the case where a set of student level controls are added to the analysis. It is also important to note that the shape suggested by Figure 1 could arise from a deficiency in test design in which the scores are not comparable over time. Since the math exams used in our application were designed to be comparable across grades, we trust that this problem will not be significant in our analysis.

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

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
$\Pr(\cdot > 0 \vert D)$	.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 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

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

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 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.

4.3 Do Schools in the Philippines Matter?

The results in the previous section have described a very general and important feature of our model, as they documented the differences between our approach to evaluating school effects and the approach taken by 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 in the context of our specific application if schools in the Philippines really matter.

To obtain improved, though not ideally "causal" estimates of the school effects α_s, we take up the original specification in (1) and (2) which includes a set of student-level controls in z_{is}. At this point, however, we do not include any school characteristics Q since it is premature to search for characteristics explaining differences across schools before first establishing that Philippine schools do indeed differ in the production of student achievement.

The characteristics in z_{is} include 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). Our hope is that accounting for these characteristics will mitigate (though probably not eliminate) selectivity concerns since wealthy, well-educated parents may elect to enroll their students in "better" schools, thus confounding school performance with household characteristics. When these controls are added, we interpret school effects as net of these family background measures. This strategy of attempting to reduce endogeneity concerns by including several family controls has been advocated in past work[13] and is, perhaps, the best we can do with our data given the lack of an obvious exclusion restriction or natural experiment.

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_ϵ^2	124.24	(3.62)	124.89	(3.87)	125.74	(3.80)
σ_α^2	5.73	(1.62)	5.93	(1.85)	7.04	(2.28)
$\sigma_\alpha^2/(\sigma_\alpha^2+\sigma_\epsilon^2)$	.044	(.012)	.045	(.013)	.053	(.017)

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

[13]See, e.g., Dearden, Ferri and Meghir (2002). They write (page 3) "In our view, the best way to deal with the endogeneity issues with data such as ours is to control for the variables that are likely to be driving school selection before the relevant treatment occurs." Their controls, like ours, include an "ability" measure (which in our case would be the intake test) and family characteristics.

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: $\sigma^2_\alpha/(\sigma^2_\alpha + \sigma^2_\epsilon)$. Across all of our samples, we find between 4.4 to 5.3 percent of the 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 $\sigma^2_\alpha/(\sigma^2_\alpha + \sigma^2_\epsilon)$ was ever less than .019, suggesting that school-level variation accounts for at least 2 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.

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. 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 8 point increase in test scores (approximately one-half of a standard deviation in 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? An Intuitive Test

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 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 small 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)).

Reallocating Students to "Pseudo-Schools"

To address these concerns we consider an intuitive test and apply it our data. In this test we keep the vector of school ID's fixed, thus preserving the number of schools (72) as well as the distribution of observations per school that is present in our data. We then create "pseudo-schools" by randomly re-ordering our observations and assigning the reshuffled observations to the original school ID vector. This reallocation keeps the (x, y) pairs (i.e., the intake and follow-up scores) intact and will not affect the total error variance $(\sigma_\epsilon^2 + \sigma_\alpha^2)$. The *fraction of the total variance* attributable to the "schools," 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 .3, and no trial produced an estimate of σ_α^2 remotely close to the estimate we obtained using our original data, which was 5.93. This suggests that there are indeed important school effects in our data, as no artificial reorganization of individuals to like groups produces comparable evidence of group effects.

4.4 Adding Measures of School Quality

Given strong evidence that schools in the Philippines do indeed differ, we move on to find school characteristics (Q_s) that may explain variation in school performance.[14] To be specific, 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) in the regression equation in (2).[15] All of the quality variables are standardized to have mean zero and unit variance.

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	$\Pr(\cdot > 0 \mid D)$	Post Mean	Post Mean	$\Pr(\cdot > 0 \mid D)$	Post Mean	Post Mean	$\Pr(\cdot > 0 \mid D)$
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_ϵ^2	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_α^2	5.30	(2.08)	1.00	3.89	(1.72)	1.00	5.22	(1.92)	1.00
$\frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma_\epsilon^2}$	.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.

[14] We do not interpret the parameters which follow as strictly causal estimates, but instead characterize them as the estimated impacts of various school characteristics once a reasonably rich set of family controls and initial student achievement have been included in the analysis.

[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 (number of books in the library per student) was not significant in most of the specifications with controls for Electricity.

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.

In terms of the school quality variables Q, we first note that the electricity coefficient is strongly positive across all three of our samples. Specifically, schools with electricity tend to outperform schools without it by about 4.28-4.90 points across our various samples.[16] 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. If we look 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) did 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

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

but characteristics of the teachers and student-teacher ratios play relatively little role in explaining variation in school performance.

It is also worthwhile to investigate how the estimated impacts of the school quality variables change as the treatment of the outcome variable changes (e.g., adopting a value-added, linear or semiparametric model). To this end, we report in Table 5 point estimates (posterior means) of the coefficients associated with variables in z and Q across the three alternate model specifications.

Table 5
Coefficient Posterior Means from Alternate Models
Value-Added (VA), Linear (LIN) and Semiparametric (SEM)

	Sample Restriction $n_s \geq 3$			Sample Restriction $n_s \geq 5$			Sample Restriction $n_s \geq 10$		
	VA	LIN	SEM	VA	LIN	SEM	VA	LIN	SEM
X Variables									
Constant	7.67	6.89	—-	7.71	7.15	—-	7.30	6.70	—-
Intake Score	—-	.567	—-	—-	.564	—-	—-	.552	—-
Male	-2.38	-3.03	-3.22	-2.25	-3.01	-3.24	-2.34	-3.18	-3.41
MomEd	-.103	.088	.120	-.089	.098	.117	-.082	.091	.102
DadEd	.029	.176	.167	.037	.175	.177	.061	.194	.202
LogExpend	.370	1.09	.765	.352	1.07	.829	.358	1.09	.988
Q Variables									
ClassSize	-.113	.192	.468	-.051	.183	.119	-.127	-.084	-.005
InService	.153	.121	.057	.120	.067	.102	.275	.139	.076
TeachExp	-1.01	-1.74	-2.02	-1.15	-2.07	-2.18	-1.16	-1.99	-2.11
Electricity	.715	1.34	1.54	.846	1.74	1.91	.814	1.85	2.01

As a general rule, predictions from the linear model remain reasonably close to those obtained from the semiparametric hierarchical model. This is to be expected given the similarity in school effects that was documented in section 4.2 and Figure 2. The value added predictions, however, can be quite different. First, note that the slope coefficient on the intake test score in the linear model is approximately .56 (with a small posterior standard deviation), indicating clear evidence against the value-added model which imposes a unit slope coefficient. The magnitudes of the school quality coefficients associated with teacher experience and electricity (i.e., the "significant" school characteristics) in the value-added model are essentially one-half of those obtained from the linear and semiparametric models. This large discrepancy across coefficients translates into meaningful differences in predictions regarding the effectiveness of school resources in the Philippines. Finally, coefficients associated with some first-stage parameters in z (such as mother's education) are often of different signs than those obtained in the linear and semiparametric models. When properly accounting for the relationship between intake and follow-up scores, we obtain larger and seemingly more sensible impacts of parental education on student achievement. The overall sensitivity of key coefficients to the specification of the test score relationship suggests that researchers should be concerned about functional form issues in addition to typical endogeneity considerations.

5 Conclusion

Using a rich set of data from Cebu, Philippines, we introduce and estimate a semiparametric hierarchical model of achievement growth that nests the widely-used value-added and linear models. We find evidence against both specifications and also find that value-added models give significantly different predictions regarding school performance, school rankings, and the effectiveness of school inputs than either the linear or semiparametric models.

Using our hierarchical specification and conducting some intuitive statistical tests, we find strong evidence that schools in the Philippines matter in the production of student achievement. Our results regarding the effectiveness of school inputs are also 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.

Although we do find that schools clearly matter, only 4.4 to 5.3 percent of the conditional variation in follow-up scores can be explained by variation in performances 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

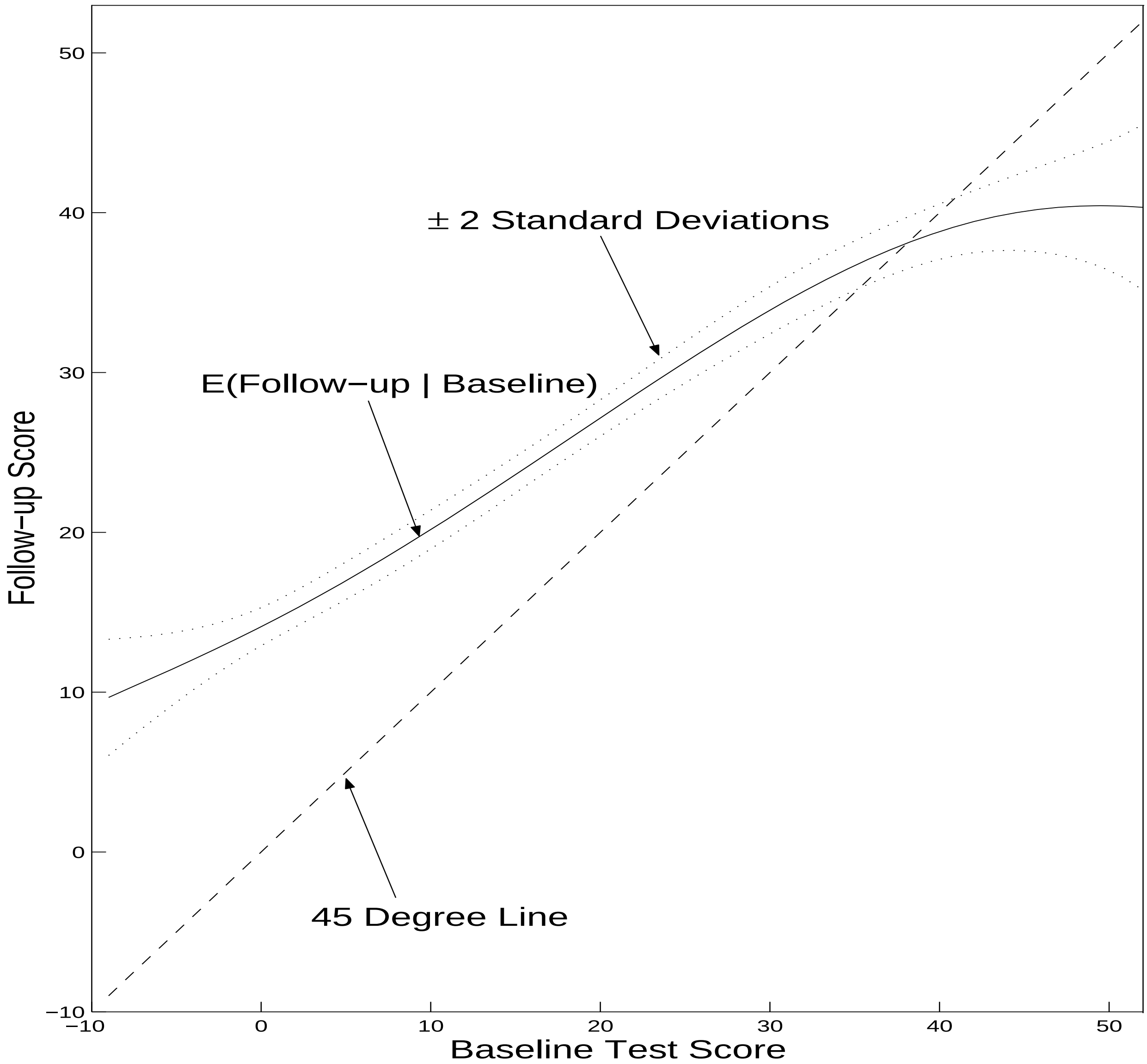
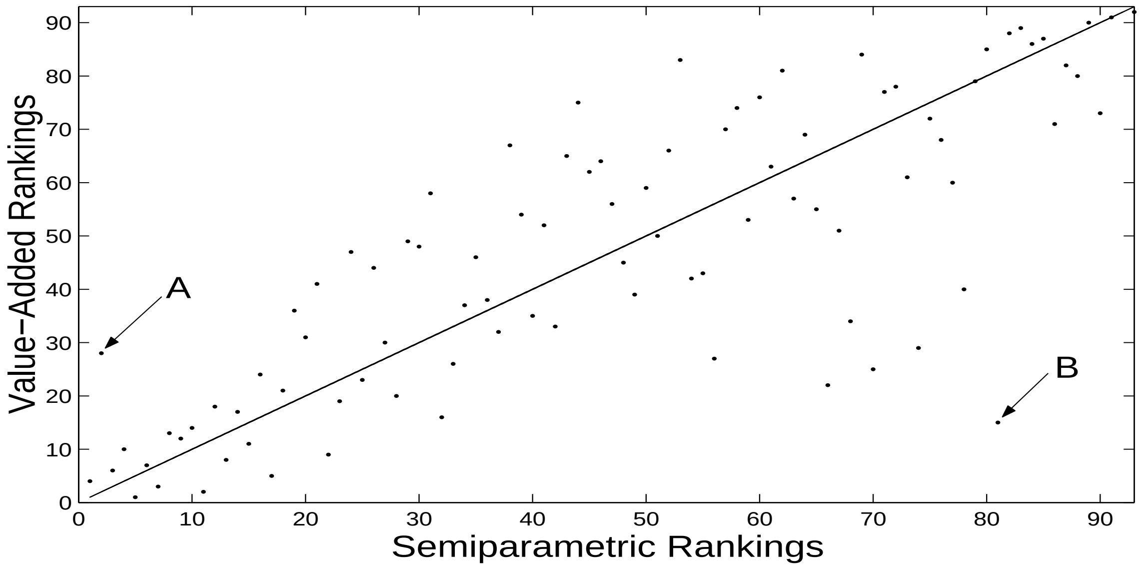


Figure 2: Comparison of Point Estimates of School Rankings Across Various Models



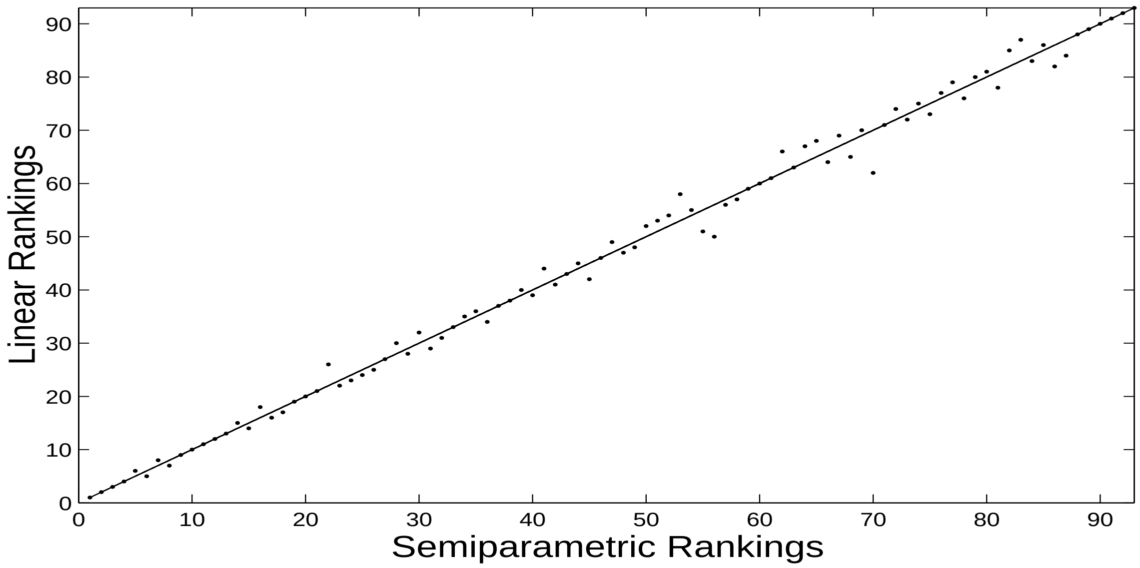


Figure 3: Boxplot of 72 School-Specific Random Effect Posteriors ($p(\alpha_s|\text{Data})$), Ordered by their Posterior Medians

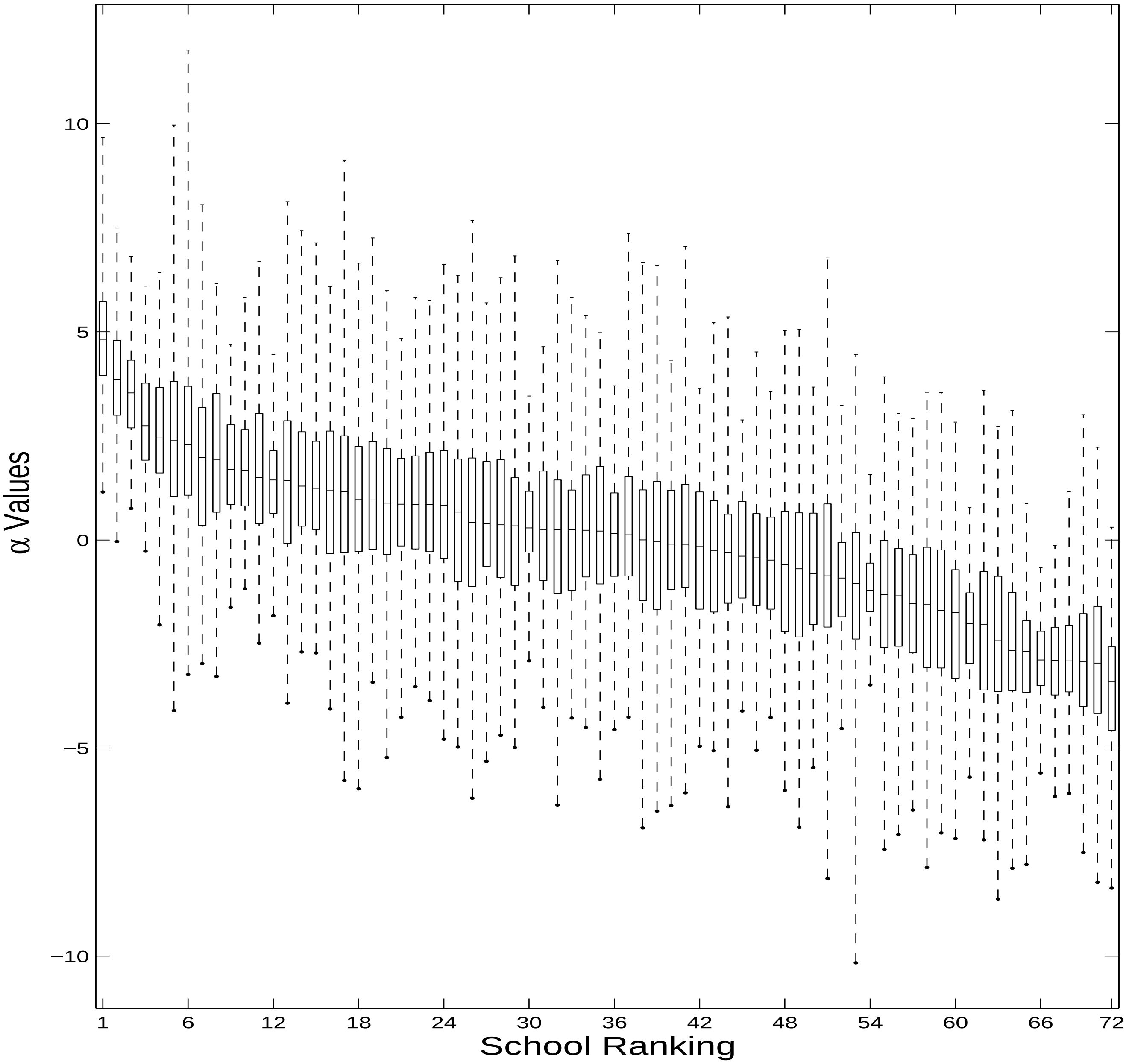
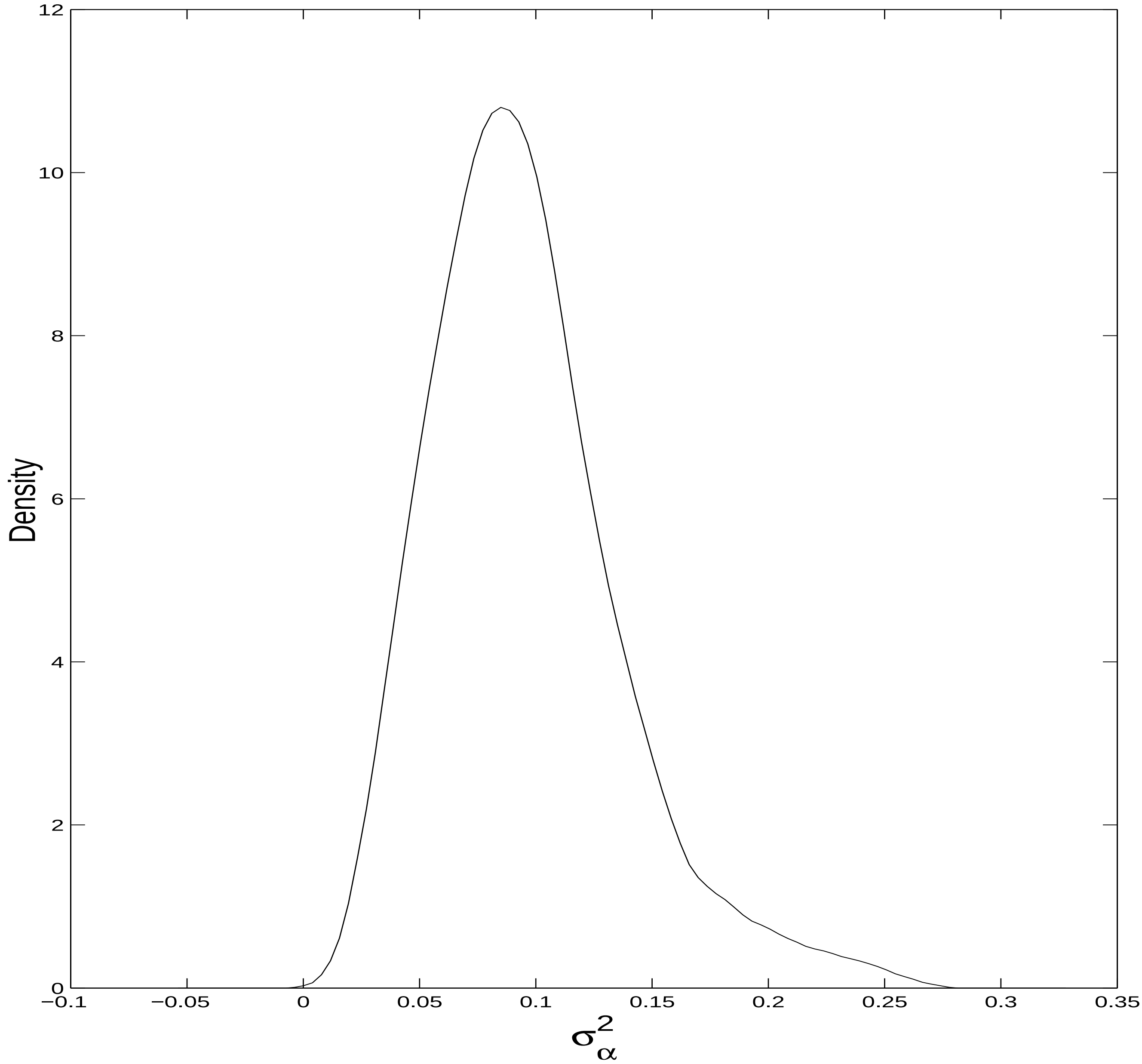


Figure 4: Density of $\hat{\sigma}_{\alpha}^2$ values obtained by Randomly Reallocating Individuals to "Pseudo=Schools"



7 Appendix B: The Posterior Simulator

We employ a Bayesian estimation approach and fit the model in (1) and (2) using the *Gibbs sampler*. As in Smith and Kohn (1996), we represent f as 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 $\tau_j < \tau_{j+1}$.

Let $X_{n\times(J+1)} = [\mathbf{1}\ x\ x^2\ (x-\tau_1)_+^3\ \cdots\ (x-\tau_{J-2})_+^3]$. We 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. In practice we allow for 12 potential knot points and equally space them throughout the baseline test score support $[-8, 53]$.

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 $c = n\ \pi_j = 1/2\ \forall j$, $\mu_\beta = \mu_\gamma = 0$, $a = 3$, $b = 1/2$ and choose V_β and V_γ as diagonal matrices with diagonal elements equal to 100.

The Blocked Gibbs Algorithm

The parameters of our model consist of $\Gamma = [\theta\ \delta_\theta\ \sigma_\epsilon^2\ \{\alpha_s\}\ \gamma\ \sigma_\alpha^2]$. To obtain draws from this posterior, we employ the Gibbs sampler and cycle though the following blocked posterior conditionals described below. 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 this joint conditional posterior by sampling from the following densities:

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

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

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

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

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

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

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.

8 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. Working paper, Department of Economics, University of California-Irvine.

[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] Dearden, L., J. Ferri and C. Meghir. 2002. The Effect of School Quality on Educational Attainment and Wages. *Review of Economics and Statistics* 84(1), 1-20.

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

[9] 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.

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

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

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

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

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

[15] 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.

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

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

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

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

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

[21] 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.

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

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

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

[25] Koop, G. and J.L. Tobias. 2005. Semiparametric Bayesian Inference in Smooth Coefficient Models, *Journal of Econometrics*, forthcoming.

[26] 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.

[27] 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.

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

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

[30] Tobias, J.L. 2004. Assessing Assessments of School Performance: The Case of California, *The American Statistician* 58(1), 2004.