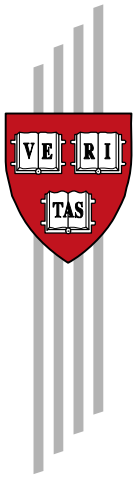
Do Value-Added Estimates Add Value? Accounting for Learning Dynamics

Tahir Andrabi, Jishnu Das, Asim I. Khwaja, and Tristan Zajonc

CID Working Paper No. 158
March 2008



Working Papers

Center for International Development at Harvard University

Do Value-Added Estimates Add Value? Accounting for Learning Dynamics*

Tahir Andrabi — Pomona College

Jishnu Das — World Bank

Asim I. Khwaja — Harvard University

Tristan Zajonc — Harvard University

First Draft: April 18, 2007. This Draft: February 19, 2008

Abstract

Value-added estimates—estimates based on the evolution rather than level of achievement— are viewed by most researchers as more reliable than cross-sectional comparisons since they ostensibly "difference out" the influence of omitted fixed inputs, such as wealth and ability. We show that the restricted value-added model, which assumes that past achievement carries over with no loss, is clearly rejected by the data, and that the more flexible lagged value-added model is biased by measurement error and omitted heterogeneity that enters each period. Using dynamic panel methods that address these biases and data on public and private schools in Pakistan, we find that the restricted value-added model yields wildly biased estimates for the private school effect, sometimes even flipping the sign. The lagged value-added model performs better due to countervailing measurement error and heterogeneity biases. More generally, rapid and potentially heterogenous achievement decay or "fade-out", which evidence suggests underlies our results, has broad implications for experimental and non-experimental program evaluation, and value-added accountability systems.

Keywords: education, value-added model, Pakistan, panel data methods
JEL Codes: I2, C52, C23

* *Some figures may display incorrectly without Acrobat Reader 8.*

We are grateful to Alberto Abadie, Chris Avery, Pascaline Dupas, Brian Jacob, Dale Jorgenson, Karthik Muralidharan, Jesse Rothstein and seminar participants for helpful comments on drafts of this paper.

> Left to itself every mental content gradually loses its capacity for being revived, or at least suffers loss in this regard under the influence of time. Facts crammed at examination time soon vanish, if they were not sufficiently grounded by other study and later subjected to a sufficient review. But even a thing so early and deeply founded as one's mother tongue is noticeably impaired if not used for several years.
>
> —Hermann Ebbinghaus in *Memory*, 1885

1 Introduction

One of the major puzzles in education research is the so-called failure of input-based policies. In a review of 376 education production function estimates, Hanushek (2003) finds that the vast majority report negative or statistically insignificant estimates for inputs such as the teacher-pupil ratio, teacher education, teacher experience, teacher salary, education expenditure, facilities, administration, and teacher test scores. Of these many results, value-added estimates—estimates based on learning trajectories rather than levels—are deemed the "highest quality."[1] If anything, these higher quality estimates suggest input-based policies are even more ineffective: of the 41 estimates of the value-added effect of teacher education, for example, 10 percent are negative and 90 percent are statistically insignificant.

Given the popularity of value-added specifications for program evaluation (e.g. Boardman and Murnane, 1979; Hanushek, 1979; Todd and Wolpin, 2003; Hanushek, 2003) and increasing advocacy of value-added accountability systems (e.g. Doran and Izumi, 2004; McCaffrey, 2004; Gordon, Kane and Staiger, 2006), understanding if low value-added can be consistent with effective schooling is critical. We argue it can be, at least in so far as value-added models are commonly specified and estimated. The key is properly accounting for learning dynamics. We show these dynamics dramatically alter value-added estimates and have broad implications for both experimental and non-experimental program evaluation.

For a glimpse of the primary issue involved, consider the following two figures. Figure 1 illustrates what appears to be a common finding in the literature. Here, we plot test scores for 12,000 children in public and private schools in Pakistan. Across the three years shown, children in private schools significantly outperform children in public schools. At the same time, children's learning *trajectories*

[1] More precisely, Hanushek (2003, p. 79) gives this distinction to value-added estimates from a *single* state.

lie parallel. Over any two years, achievement gains are equal in public and private schools. How should we interpret this finding?

A value-added "gain-score" interpretation argues that public and private schools contribute equally to achievement. The only difference is an endowment, ostensibly attributable to omitted pupil characteristics and selection. In language common to the literature, the value-added model "differences out" this endowment and isolates the contribution of school type by looking at the trajectory rather than level of achievement.

But Figure 2 shows why interpretation is misleading. Here, we again plot the test score gain for children in public and private schools, but condition the gain on original achievement. The striking conclusion from Figure 1 is reversed. Children in private schools *gain more* than children in public school at *every* point of the initial score distribution. Because (a) gains decrease with higher initial scores and (b) children in private school start off at higher levels, the learning trajectories lie parallel even though private schools "add more value" for comparable children. While one explanation for this pattern is mean reversion due to measurement error, we show later that this is only part of the issue.

Figures 1 and 2 are graphical representations of two popular value-added specifications. Figure 1 is the analog of the restricted value-added or gain-score model. This model assumes, among other things, that lagged achievement contributes cumulatively *without loss* to future achievement. The effect of a teacher who spurs a child a few months ahead, say, will remain unabated in third, fourth, and even twelfth grade. By comparison, Figure 2 represents the more flexible lagged value-added model. Here, the contribution of previous achievement may decay. A child may forget a portion of what she learned last year.

Numerous empirical estimates of the lagged value-added model suggest a depreciation coefficient—the coefficient on lagged achievement—around 0.6 (e.g. Schwartz and Zabel, 2005; Sass, 2006; Todd and Wolpin, 2006). This is far below the rate assumed by the restricted value-added model and suggests that a specification analogous to Figure 2 is the right way to proceed. Yet the restricted specification remains popular and the lagged specification is rarely estimated in the dynamic panel framework it requires. This would not be major concern if bias or misspecification of the depreciation coefficient left the value-added estimate of the input unbiased. But it does not. If the true contribution of lagged achievement is only 0.6—we argue that even this estimate may be too

high—then differencing induces a negative correlation between contemporaneous inputs and lagged inputs, assuming inputs are serially correlated. This biases the value-added estimate downward—to zero in Figure 1.

The strong functional form assumptions imposed on the technology of skill formation by the restricted value-added model, and even the lagged value-added model, are well-known (see, for example, Boardman and Murnane, 1979; Todd and Wolpin, 2003). Less known, and certainly less addressed, are the estimation difficulties associated with including a lagged dependent variable as an explanatory variable.[2] This paper explores various dynamic panel approaches to estimating the education production function in its value-added form. These methods have received little attention in the education literature even though the lagged value-added model is almost an archetypal example of the widely studied autoregressive error components model (Hausman and Taylor, 1981; Anderson and Hsiao, 1981; Arellano and Bond, 1991; Arellano and Bover, 1995).

We highlight a simple learning dynamic: the coefficient on lagged achievement. There are two primary difficulties. First, any cumulative individual growth heterogeneity that speeds learning (rather than entering as a one-time effect) biases the estimated coefficient on lagged achievement upward. A talented child, for example, may not only enter first grade ahead but may also learn quicker. Since this talent enters in each period, lagged achievement is clearly endogenous. The input coefficients may be biased indirectly by the endogeneity of lagged achievement or directly by correlation with the omitted heterogeneity. Second, as highlighted by Kane and Staiger (2002) and Chay, McEwan and Urquiola (2005) test score measurement error is a major concern in educational datasets. In the lagged value-added model, measurement error counteracts the upward heterogeneity bias by attenuating the coefficient on lagged achievement. Whether the two effects cancel depends on nature of the omitted heterogeneity, the particular exam, and the included controls.

Following the dynamic panel literature, we survey a range of possible moment restrictions based on assumptions regarding the data generating process. The most common approach uses the twice-lagged scores as an instrumental variable in a differenced model (Arellano and Bond, 1991). However, differencing is problematic for estimating the private school effect, and many other effects common

[2]In addition to reviewing many other approaches to estimating education production functions, Todd and Wolpin (2003) mention within the context of value-added models all three difficulties we confront: misspecification, measurement error and growth heterogeneity. They do not, however, discuss the dynamic panel estimators we apply or the magnitude of the resulting biases using real data.

in education, since there is little time-series input variation—only five percent of children switch schools each year—and differencing amplifies any measurement error. We therefore also explore versions of the Arellano and Bover (1995) "system" GMM estimator that accounts for unobserved heterogeneity but allows time-invariant inputs that might otherwise be unidentifiable using more common difference GMM methods. To address measurement error, we attempt to avoid instrumental variable strategies by using analytical corrections based on the known heteroscedastic reliability ratio derived from Item Response Theory (IRT).[3]

Only a handful of papers in education have experimented with dynamic panel estimators for the lagged value-added model. In a paper developed concurrently to ours, Rothstein (2007) estimates value-added teacher effects using Chamberlain's correlated random effects model. He emphasizes the dynamic vs. static tracking of teachers and proposes a test for whether teacher assignment is strictly exogenous conditional on a student fixed effect, which he rejects.[4] Schwartz and Zabel (2005) explore various estimators of school efficiency, including one based on a dynamic panel specification. Unlike this paper, they use school-level data and seek to difference out a school-grade fixed effect. Their paper is notable, however, in its careful discussion of omitted heterogeneity, measurement error, and endogeneity.[5] Sass (2006) estimates the contribution of charter schools in Florida using the Arellano and Bond (1991) differences GMM approach. In the simplest specification, Sass's estimate is analogous to the switching estimator we discuss but does not address the measurement error that remains in the differenced model; this may partly explain the low depreciation coefficient found (~ 0.2). Santibanez (2006) also uses the Arellano and Bond (1991) estimator to analyze the impact of teacher quality on child achievement, but does not correct for measurement error. Our estimators extend these efforts and emphasizes the importance of decay.

How much does this all matter? Using data from a panel of 12,000 children in 800 public

[3]Item Response Theory is a set of statistical techniques that seek to recover the latent trait driving the response process on an exam. IRT models the behavior of each item—i.e. its difficulty, ability to discriminate between two children, and likelihood of being guessed—so that any differences in items can be removed from the score. IRT scores have cardinal meaning and the associated standard error can be used to correct for measurement error in subsequent analyses.

[4]Rothstein's test of dynamic vs. static tracking is similar to our discussion of strictly exogenous vs. predetermined inputs but the overall emphasis of the paper is quite distinct. In particular, we highlight the depreciation coefficient, which we find is crucial to identifying the private school effect, and survey a broad range of potential GMM estimators for lagged value-added models. Consistent with his goal of testing the the nature of teacher assignments, Rothstein's main specification is a value-added model that includes the full set of lagged (and potentially future) inputs. We focus on value-added models that use lagged achievement as a sufficient statistic for past inputs and assume geometric decay.

[5]Gronberg, Jansen and Naufal (2006) applies the same methods to data on public schools in Texas.

and private schools of rural Pakistan, we strongly reject the restricted value-added model. Our best estimates suggest that the coefficient on lagged achievement is around 0.5 and potentially significantly lower for math. We provide some preliminary evidence that decay cannot be explained by households' adjusting inputs to achievement shocks or teachers' targeting poorly performing students. There is also some evidence that decay rates may be heterogeneous, with faster decay associated with faster learners.

Somewhat surprisingly, naive estimates by OLS of the *lagged* model produce roughly equivalent numbers: 0.52 to 0.58. This agreement between our dynamic panel estimates and lagged OLS estimates stems from the countervailing measurement error and heterogeneity biases. Correcting only for measurement error attenuation yields estimates between 0.7 and 0.79, indicating that omitted heterogeneity cannot be ignored (since the dynamic panel estimates are lower). It also suggests that correcting for measurement error only may be worse than doing nothing.[6] That said, there is no reason to expect the biases to cancel exactly or for unobserved growth heterogeneity to be uncorrelated with the observed inputs. *In other words, value-added models require dynamic panel methods that account for both omitted growth heterogeneity and measurement error.*

The low coefficient on lagged achievement has several important implications. First, value-added estimates, which rely on an unbiased estimate of the depreciation parameter and assume all omitted heterogeneity is differenced out, can severely underestimate the relative contribution of schools where children's baseline test scores are high. We find that the restricted value-added model, for example, yields *negative* or insignificant estimates for the contribution of private schools. By comparison, the lagged value-added model, which performs better by reducing the bias in the depreciation coefficient, yields positive and statistically significant value-added estimates between 0.26 and 0.31 standard deviations a year. These estimates are consistent with some of our dynamic panel estimates, but there is no reason to believe this will always hold. The assumptions required to estimate both the restricted and lagged value-added models by OLS are strongly rejected by the data. *Contrary to conventional wisdom, (gain score) value-added models can provide worse inference than cross-sectional comparisons, and low value-added can be consistent with effective schooling.*

[6]For example, Ladd and Walsh (2002) correct for measurement error in the lagged value-added model of school effects by instrumenting using double-lagged test scores but don't address potential omitted heterogeneity. They show this correction significantly changes school rankings and benefits poorly performing districts. Our analysis suggests that this correction may do more harm than good if omitted heterogeneity biases the coefficient of lagged achievement upward.

Second, rapid and potentially heterogeneous decay makes short-run analysis inherently unreliable, even if effects are precisely estimated using experimental data. For example, Krueger and Whitmore (2001), Angrist et al. (2002), Krueger (2003), and Gordon, Kane and Staiger (2006) calculate the economic return of various educational interventions by citing research linking test scores to earnings of young adults (e.g. Murnane, Willett and Levy, 1995; Neal and Johnson, 1996). But our estimates suggest, and experimental evidence of fade-out seems to confirm, that program effects measured by cognitive test scores will fade rapidly and at different rates. While other effects may persist, there is rarely enough evidence to responsibly make the necessary projections. This applies equally to comparing interventions and ranking the performance of teachers and schools. *Short-run effects, even if experimentally estimated, are poor proxies for long-run effects and assuming no decay can vastly overstate a program's internal rate of return.*

Finally, while accounting for achievement decay is central to value-added models, it is hardly the only learning dynamic. In research complementary to ours, Todd and Wolpin (2006) find that lagged test scores may be insufficient statistics for past educational inputs; lagged inputs predict current achievement even after controlling for past achievement. Su (2004), Cunha, Heckman and Schennach (2006), Cunha and Heckman (2007*b*), and Cunha and Heckman (2007*a*) support this view. They find that education production functions are neither separable in inputs or time; cognitive and non-cognitive skills, and household and school inputs raise present *and* future learning. Further afield, memory research in psychology and neuroscience suggests that forgetting may be nongeometric and complex (Wixted and Ebbesen, 1991; Rubin and Wenzel, 1996). Indeed, short-term cognitive gains may be beyond the point: Currie and Thomas (1995), Garces, Thomas and Currie (2002) and Deming (2007) all find long term effects of Head Start even though test scores gains fade quickly. *Skill formation is a far richer process than the linear accumulation of knowledge.*

The rest of the paper is organized as follows. Section 2 presents the basic education production function analogy and discusses the specification of the value-added model. Section 3 emphasizes the importance of specification bias, measurement error and omitted heterogeneity, and surveys the applicability of dynamic panel methods. Section 4 summarizes our data on public and private school in Pakistan. Section 5 reports our main results and several robustness checks. Section 6 discusses alternate interpretations of decay and highlights several implications for both experimental and non-experimental program evaluation. Section 7 concludes.

2 Learning Framework

Empirical estimates of the "education production function" (EPF) generally begin with the idea that current achievement is a function of all previous inputs. Boardman and Murnane (1979) and Todd and Wolpin (2003) provide two excellent accounts of this approach and the assumptions it requires. There are many possible formulations. The data can include single or multiple cohorts of children, and inputs (and fixed effects) can be broken down into a child, household, classroom, school, and state hierarchy. These inputs can interact in any number of ways, but researchers generally assume the model is additively separable across time, and that input interactions can be captured by separable linear interactions.[7]

To avoid unnecessary clutter, and to keep the notation consistent with the dynamic panel literature, we lump all inputs into a single vector $\mathbf{x}_{it}$ and consider a single cohort. The input vector may have a hierarchical structure, as often emphasized in the education literature, and may include contemporaneous interactions. We exclude interactions between past and present inputs. Achievement for child i at time (grade) t is therefore

$$y_{it}^{*} = \boldsymbol{\alpha}_1' \mathbf{x}_{it} + \boldsymbol{\alpha}_2' \mathbf{x}_{i,t-1} + ... + \boldsymbol{\alpha}_t' \mathbf{x}_{i1} + \sum_{s=1}^{s=t} \theta_{t+1-s} \mu_{is}, \tag{1}$$

where y_i^* is true achievement, measured without error, and the summed μ_{is} are cumulative productivity shocks. Here, the input (and shock) coefficients are unique to each lag, so that the contribution of inputs to current achievement can change with time. An input applied in first grade need not have the same effect on third grade achievement as on twelfth grade achievement. That is, an input's contribution to achievement can decay (or grow) with time. Note that we not allow inputs' effects to change with age. As written, the immediate impact of private schools in third grade is the same as the immediate impact of private schools in fourth grade. While potentially unrealistic, we do not review more flexible learning models because Todd and Wolpin (2003) provide an excellent overview already and because our focus is on value-added models that use lagged achievement as a sufficient statistic for past inputs.

Estimating (1) is generally impossible because researchers never observe the full set of inputs,

[7] Cunha, Heckman and Schennach (2006) and Cunha and Heckman (2007*a*) are a notable exceptions to this pattern. They estimate a nonlinear CES education production function with dynamic complementarity between early and late investments and between cognitive and non-cognitive skills.

past and present. The value-added strategy makes estimation feasible by rewriting (1) to avoid the need for past inputs. Applying Koyk transformation—adding and subtracting βy_{it}—yields

$$\begin{aligned} y^*_{it} &= \boldsymbol{\alpha}_1' \mathbf{x}_{it} + \beta y^*_{i,t-1} + \mu_{it} && (2) \\ &\quad + (\boldsymbol{\alpha}_2' - \beta \boldsymbol{\alpha}_1') \mathbf{x}_{i,t-1} + ... + (\boldsymbol{\alpha}_t' - \beta \boldsymbol{\alpha}_{t-1}') \mathbf{x}_{i1} && (3) \\ &\quad + \sum_{s=1}^{s=t-1} (\theta_{t+1-s} - \beta \theta_{t-s}) \mu_{is}, && (4) \end{aligned}$$

where we have normalized θ_1 to unity. As Boardman and Murnane (1979) and Todd and Wolpin (2003) point out, if coefficients decline geometrically ($\alpha_j = \beta \alpha_{j-1}$ and $\theta_j = \beta \theta_{j-1}$ for all j), then previous inputs and shocks drop away and we have the familiar *lagged-value-added model*

$$y^*_{it} = \boldsymbol{\alpha}_1' \mathbf{x}_{it} + \beta y^*_{i,t-1} + \mu_{it} \tag{5}$$

The basic hope of this specification is that lagged achievement will capture the contribution of all previous inputs and any past unobservable endowments or shocks. We call α the *input coefficient* and β the *depreciation coefficient* ($\beta = 1 - \delta$). Finally, if we assume further that $\beta = 1$ we are left with the "gain score" or *restricted value-added model*

$$y^*_{it} - y^*_{i,t-1} = \boldsymbol{\alpha}_1' \mathbf{x}_{it} + \mu_{it}. \tag{6}$$

This model asserts that achievement gains do not decay with time, or, equivalently, that an input's effect on current achievement does not depend on when it was applied.

Deriving the value-added model from an cumulative education production function with the potential for depreciation is not the only starting point. Ben-Porath (1967) and Cunha and Heckman (2007*b*), for example, include lagged achievement directly in the education production function. In this learning framework, the lagged value-added model (5) is a log-linearization of the Cobb-Douglas production function

$$\exp(y^*_{it}) = \prod_k \left[\exp(x_{k,it})^{\alpha_{1k}}\right] \exp(y^*_{i,t-1})^{\beta} \exp(\mu_{it})$$

where k denotes different inputs. In the language of Cunha and Heckman (2007*b*), β captures the

self-productivity of past achievement. This formulation also highlights an alternative interpretation of coefficient on lagged achievement: concavity. For consistency we generally use the term depreciation coefficient for β. We discuss alternative interpretations in more depth in Section 6. For now it is enough to note that none of the estimation issues depend on one's interpretation.

3 Estimating Value-Added Learning Models

3.1 Misspecification, Measurement Error and Growth Heterogeneity Bias

There are two primary difficulties in estimating (5). First, the error term μ_{it} may include individual growth heterogeneity (e.g. $\mu_{it} \equiv \eta_i + \upsilon_{it}$). Lagged achievement only captures individual heterogeneity if it enters through a one-time process. A child who enters first grade already reading, for example, has an one-time endowment that will be captured by lagged achievement. At the same time, this child may be unusually talented or privileged and may learn faster. Since this unobserved heterogeneity enters in each period, $\text{Cov}(y^*_{i,t-1}, \mu_{it}) > 0$ and β will be biased. Second, test scores are inherently a noisy measure of latent achievement. Letting $y_{it} = y^*_{it} + \varepsilon_{it}$ denote some measure of achievement, we can rewrite the latent lagged value-added model (5) in terms of observables. The full error term now includes measurement error, $\mu_{it} + \varepsilon_{it} - \beta\varepsilon_{i,t-1}$. Dropping all the inputs to focus solely on the depreciation coefficient, the expected bias is

$$\text{plim}\, \beta_{OLS} = \beta + \left(\frac{\text{Cov}(\eta_i, y^*_{i,t-1})}{\sigma^2_{y^*} + \sigma^2_\varepsilon} \right) - \left(\frac{\sigma^2_\varepsilon}{\sigma^2_{y^*} + \sigma^2_\varepsilon} \right) \beta. \tag{7}$$

We can see that the coefficient is biased upward by growth heterogeneity and downward by measurement error. These effects only cancel exactly when $\text{Cov}(\eta_i, y^*_{i,t-1}) = \sigma^2_\varepsilon \beta$ (Arellano, 2003).

To the degree β is biased, so too will the input coefficient $\boldsymbol{\alpha}$. Generally, the input coefficient will be biased downward if $\hat{\beta} > \beta$ and upward otherwise. This is because upward bias in the depreciation coefficient means we are actually estimating

$$y_{it} - \hat{\beta} y_{i,t-1} = \boldsymbol{\alpha}' \mathbf{x}_{it} + [(\beta - \hat{\beta}) y_{i,t-1} + \mu_{it} + \varepsilon_{it} - \beta\varepsilon_{i,t-1}].$$

Imposing the biased estimates illustrates how the resulting error term now includes $(\beta - \hat{\beta}) y_{i,t-1}$.

Since inputs and lagged achievement are positively correlated, the input coefficient will, in general, by biased downward if $\hat{\beta} > \beta$. The precise bias, however, depends on the degree of serial correlation of inputs and the potential correlation between inputs and growth heterogeneity that remains in μ_{it}.

The restricted value-added model avoids some of these issues by assuming the lagged achievement does not depreciate, that is $\beta = 1$. An input in first grade must have the same effect on third grade achievement as on twelfth grade achievement. The model then estimates

$$y_{it} - y_{i,t-1} = \alpha' \mathbf{x}_{it} + [(\beta - 1)y_{i,t-1} + \mu_{it} + \varepsilon_{it} - \beta\varepsilon_{i,t-1}], \tag{8}$$

where the brackets contain the full error term. If $\beta = 1$, as assumed, and inputs are uncorrelated with μ_{it}, OLS yields consistent estimates of the parameters. But widespread evidence suggests that $\beta < 1$ (e.g. Schwartz and Zabel, 2005; Sass, 2006; Todd and Wolpin, 2006) . In this case, OLS yields

$$\operatorname{plim} \hat{\alpha}_{OLS} = \alpha - (1 - \beta)\frac{\operatorname{Cov}(\mathbf{x}_{it}, y_{i,t-1})}{\operatorname{Var}(\mathbf{x}_{it})} + \frac{\operatorname{Cov}(\mathbf{x}_{it}, \eta_i)}{\operatorname{Var}(\mathbf{x}_{it})}. \tag{9}$$

There are two competing biases. By assuming an incorrect depreciation coefficient we leave a portion of past achievement in the error term. This misspecification biases the input coefficient downward by the first term in (9). The second term captures possible correlation between current inputs and omitted growth heterogeneity. If there is none, as the restricted value-added model assumes, then the second term is zero and the bias will be unambiguously negative.

External instruments and even randomization do not overcome this problem entirely. Many instruments operate over multiple years. If children were randomized into schools in first grade we eliminate the second bias—the correlation between inputs and the omitted effect—but leave the first unchanged since $\operatorname{Cov}(\mathbf{z}_{it}, y_{i,t-1}) \neq 0$ if the instrument operates on lagged input choices (school type in grades one and two). The bias becomes unambiguously negative. While we can estimate the two-year effect, we cannot compare it to one year estimates by dividing by two, as the model implies. *Indeed, it is impossible to compare interventions of different lengths without a consistent estimate of the depreciation coefficient β.*

3.2 Addressing Measurement Error in Test Scores

As highlighted by Kane and Staiger (2002) and Chay, McEwan and Urquiola (2005), measurement error and sampling volatility dominate educational datasets. For accountability schemes, noise hides exceptional (or dismal) teachers and schools. For the lagged value-added model, measurement error attenuates the coefficient on lagged achievement and biases the input coefficient in the process. Given that test score reliability ratios generally range between 0.7 and 0.9, this bias may be quite severe.

Instrumental variables are one simple way to correct for measurement error. Lagged scores or alternate subjects are good candidates. With item level data, one might even instrument for the second half of the test using the first half. But IV strategies have a number of drawbacks. In the dynamic panel models we discuss shortly, correcting for measurement error using lags requires four years of data for each child—a burdensome requirement. Alternate subjects do not require additional years, but may contain correlated measurement error—the "barking dog" effect. Even absent these concerns, instrumenting with lags and alternate subjects can be inefficient.

Instrumenting may also identify a different parameter if correlated components of achievement behave differently than uncorrelated components (Imbens and Angrist, 1994). For example, the portion of third grade achievement that remains correlated with fourth grade achievement may decay at a different rate than what was learned most recently. This is particularly true in an optimizing model of skill formation where parents smooth away shocks to achievement. In such a model, unexpected shocks to achievement (beyond measurement error) would fade more quickly than expected gains. Instrumenting using contemporaneous alternate subject scores will therefore likely identify different parameters than instrumenting using previous year scores. While this type of heterogeneity is important, it is largely beyond the scope of this paper.

In place of instrumental variable strategies, we correct for measurement error analytically using the standard error of each test score, available from Item Response Theory.[8] Because the standard error is heteroscedastic—tests discriminate poorly between children at the tails of the ability distribution—we use the heteroscedastic errors-in-variables (HEIV) procedure outlined in Sullivan (2001) and followed by Jacob and Lefgren (2005) among others. This method is similar to the typical

[8]Item Response Theory provides the standard error for each score from the inverse Fisher information matrix after ML estimation of the IRT model (Birnbaum, 1968). This standard error is reported in many educational datasets.

errors-in-variables regression, but uses a heteroscedastic reliability ratio to construct an empirical Bayes estimate of the test score. This estimate is the best linear predictor of the true score and is uncorrelated with the measurement error by the first order conditions that define it (Whittemore, 1989; Sullivan, 2001). In the appendix we provide a more detailed explanation of this analytical correction.

3.3 Dynamic Panel Approaches to the Value-Added Model

The lagged value-added model (5) can be interpreted as a autoregressive dynamic panel model with individual effects:

$$y^*_{it} = \boldsymbol{\alpha}'\mathbf{x}_{it} + \beta y^*_{i,t-1} + \mu_{it}, \tag{10}$$

$$\mu_{it} \equiv \eta_i + \upsilon_{it}. \tag{11}$$

This econometric model has received considerable attention dating back to at least Hausman and Taylor (1981) and Anderson and Hsiao (1981).[9] Our approach follows Arellano and Bond (1991) and Arellano and Bover (1995) who propose a linear GMM estimator that, in the simplest case, uses moments from a differenced equation to identify β.

The dynamic panel literature has spawned numerous suggestions for possible moment conditions. We focus on simple linear moment conditions and split our analysis into three groups. The first, "differences GMM", is based on differencing (10) to eliminate the fixed effect. It assumes inputs are exogenous or predetermined conditional on the fixed effect and can be viewed as switching estimators—the input effect is identified from time-series variation such as children switching schools. The second, "difference and levels GMM", assumes inputs are uncorrelated or constantly correlated with the omitted effects. These system estimators are somewhat more complex because instruments are available in both the differenced equation and the levels equation. Finally, we consider a simplified "levels-only" GMM estimator which uses moments generated from a conditional stationarity assumption to fully identify the model without relying on a system of equations. While this assumption is restrictive in the context of learning, the model is simple to estimate (i.e. one can estimate it easily via univariate 2SLS) and it reduces the bias compared to OLS. Table 1 provides

[9] See Arellano and Honore (2001) and Arellano (2003) for excellent reviews of this and other panel models.

a brief summary of the estimators we explore; the sections below discuss these estimators in more depth.

3.3.1 Differences GMM: Switching estimators

The value-added model differences out omitted endowments that might be correlated with the inputs. It does not, however, difference out growth heterogeneity that speeds learning. To accomplish this, one natural idea is to difference again. Indeed, this is the basic intuition behind the Arellano and Bond (1991) difference GMM estimator.

Arellano and Bond (1991) begin by differencing—hence "differences GMM"—the dynamic panel specification of the lagged-value-added model (10) to get

$$y^*_{it} - y^*_{i,t-1} = \boldsymbol{\alpha}'(\mathbf{x}_{it} - \mathbf{x}_{i,t-1}) + \beta(y^*_{i,t-1} - y^*_{i,t-2}) + [\upsilon_{it} - \upsilon_{i,t-1}]. \tag{12}$$

Here, the differenced model eliminates the unobserved fixed effect η_i. Yet this model cannot be estimated by OLS since the $y^*_{i,t-1}$ is correlated by construction with $\upsilon_{i,t-1}$ in the error term. As a solution, Arellano and Bond (1991) propose instrumenting for $y^*_{i,t-1} - y^*_{i,t-2}$ using lags two periods and beyond, such as $y^*_{i,t-2}$ or certain inputs, depending on the exogeniety conditions. These lags are uncorrelated with the error term but correlated with the change in lagged achievement.[10]

The Arellano and Bond (1991) estimator does not address measurement error on its own. If we replace true achievement with observed achievement, (12) becomes

$$\triangle y_{it} = \boldsymbol{\alpha}'\triangle\mathbf{x}_{it} + \beta\triangle y_{i,t-1} + [\triangle\upsilon_{it} + \triangle\varepsilon_{i,t} - \beta\triangle\varepsilon_{i,t-1}]. \tag{13}$$

The potential instrument $y_{i,t-2}$ is uncorrelated with $\triangle\upsilon_{it}$ but is correlated with $\triangle\varepsilon_{i,t-1} = \varepsilon_{i,t-1} - \varepsilon_{i,t-2}$ by construction. Additional lags or alternate subjects would not necessarily be correlated with the error term and could be used as instruments. We choose to correct for measurement error by replacing the left-hand achievement variables and instrument with our empirical Bayes estimates.[11]

The implementation of the difference GMM approach depends on the precise assumptions about

[10] Using double lagged dependent variables as instruments also requires that shocks υ_{it} are serially uncorrelated.

[11] It is important in this differenced model to construct the empirical Bayes estimates using the full set of test scores. Otherwise, the instrument $\tilde{y}_{i,t-2}$ may be correlated with the residual $y^*_{i,t-1} - \tilde{y}_{i,t-1}$ in the error term, for example. See the appendix for details.

inputs. We consider two candidates: strictly exogenous inputs and predetermined inputs. Both identify the input coefficient from time-series variation and can be viewed as switching estimators. Using the superscript notation $\mathbf{x}_i^t = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{it})'$ and substituting our empirical Bayes estimates $\tilde{y}_{it}$ for unobserved achievement y_{it}^*, strict exogeneity implies:

Assumption 1 (Strictly Exogenous Inputs) *Inputs are uncorrelated with past, present, and future disturbances, and the lagged (EB) achievement is uncorrelated with present and future disturbances. That is,*

$$E[\upsilon_{it}|\mathbf{x}_i^T, \tilde{y}_i^{t-1}] = 0 \qquad (t = 2, ..., T). \tag{14}$$

Strict exogeneity does not rule out correlation between inputs and the omitted fixed effect, but provides the needed orthogonality conditions for the differenced model. In a differenced model, strict exogeneity (14) implies

$$E[\upsilon_{it} - \upsilon_{i,t-1}|\mathbf{x}_i^T, \tilde{y}_i^{t-2}] = 0 \qquad (t = 3, ..., T), \tag{15}$$

which yields $\mathbf{x}_i^T$ and $\tilde{y}_i^{t-2}$ as valid instruments. Subject to the standard rank conditions, these moment conditions identify the coefficients for time-varying inputs and the levels intercept, provided $T \geq 3$. In practice, the levels intercept is generally ignored, so the model is called differences GMM to indicate the the parameters are identified from a differenced model (12). Time-invariant inputs drop out of the estimation and are not identified.

In our application to public and private schools in Pakistan ($T = 3$), leads and lags of $\mathbf{x}_{it}$ make the model overidentified. Given large downward bias in asymptotic standard errors for two-step GMM found by Blundell and Bond (1998), we estimate the model by one-step GMM. The one-step estimator is equivalent to 2SLS with a GLS transformation to eliminate the moving average error structure in the differenced model, and is asymptotically equivalent to two-step GMM when υ_{it} are i.i.d.

Strict exogeneity assumes past disturbances do not affect current and future inputs, ruling out feedback effects. In the educational context, this is a strong assumption. A child who experiences a positive or negative shock may adjust inputs in response. In our case, a shock may cause a child to switch schools. To account for this possibility, we consider the weaker case where inputs are

predetermined but not strictly exogenous.

Assumption 2 (Predetermined Inputs) *Inputs are uncorrelated with present and future disturbances, but potentially correlated with past disturbances, and lagged (EB) achievement is uncorrelated with present and future disturbances. That is,*

$$E[\upsilon_{it}|\mathbf{x}_i^t, \tilde{y}_i^{t-1}] = 0 \qquad (t = 2, ..., T). \tag{16}$$

In the differenced model, predeterminedness implies

$$E[\upsilon_{it} - \upsilon_{i,t-1}|\mathbf{x}_i^{t-1}, \tilde{y}_i^{t-2}] = 0 \qquad (t = 3, ..., T). \tag{17}$$

The only difference compared to strict exogeneity is that we only use lagged $\mathbf{x}_i^{t-1}$ as instruments. Switching schools is instrumented by the original school type, allowing switches to depend on previous shocks. In practice, the instruments may be quite weak, a problem that we will encounter in our result section.

A simple test of strict exogeneity—emphasized in the education context by Rothstein (2007) in his analysis of static and dynamic teacher assignment—is to include future inputs in the differenced model. Under the null that inputs are strictly exogenous (conditional on the student fixed effect), the coefficient of future inputs should be zero barring input measurement error. Intuitively, future inputs should not have a causal influence on past achievement gains (Rothstein, 2007). Obvious violations of this assumption are when parents adjust inputs to previous shocks or based on expectations about future decisions.

3.3.2 Levels and Differences GMM: Uncorrelated or constantly correlated effects

A major difficulty with the differences GMM approach is that time-invariant inputs are unidentified. In education, many inputs have little or no time-series variation; only five percent of children switch schools, and no children switch gender, race, or parents. Even when inputs do vary, researchers often only observe one set of baseline characteristics.

We address these concerns using the levels and differences GMM framework proposed by Arellano and Bover (1995) and extended by Blundell and Bond (1998). We consider inputs that are

uncorrelated with the omitted effects—a leading case being inputs for which we have external instruments—and inputs that have constant correlation with the omitted effects.

Assumption 3 (Predetermined Inputs Uncorrelated with the Effects) *Inputs are uncorrelated with present and future disturbances, but potentially correlated with past disturbances; inputs are uncorrelated with the fixed effect; and lagged (EB) achievement is uncorrelated with present and future disturbances. That is,*

$$E[\upsilon_{it}|\mathbf{x}_i^t, \tilde{y}_i^{t-1}] = 0, \qquad (t = 2, ..., T) \tag{18}$$

$$E[\eta_i|\mathbf{x}_i^T] = 0 \tag{19}$$

Compared to differences GMM, we now use inputs $\mathbf{x}_i^t$ as instruments in the levels model (10). On the one hand, the required assumption is very strong; inputs are often correlated with the fixed effect. Certainly the decision to attend private school may be correlated with the fixed effect. But, at the same time, the assumption is weaker than OLS estimation of lagged value-added model requires; the omitted fixed effect may be correlated with lagged achievement. There are three common situations where this assumption is reasonable: 1) when we instrument for the input of interest using an external variable (such as early childhood randomization), 2) when we view inputs that may be correlated with the effect as proxy variables[12], and 3) when we consider η_i a random effect.

Another possibility we consider is predetermined inputs that have a constant correlation with the individual effects. Arellano and Bover (1995) present this model for predetermined variables and Bhargava and Sargan (1983) and Breusch, Mizon and Schmidt (1989) present it for strictly exogenous variables. In lagged value-added model, these estimators can be viewed as levels and differences switching estimators, since they rely on time-series variation in both the levels and differences equations.

Assumption 4 (Predetermined Inputs and Constant Correlation with Effects) *Inputs are uncorrelated with present and future disturbances, but potentially correlated with past disturbances; inputs have constant correlation with the fixed effect; and lagged (EB) achievement is uncorrelated*

[12]Note that viewing time-varying inputs as proxy variables creates difficulties in a system-GMM context because the coefficient in the differenced equation will be different than the coefficient in the levels equation.

with present and future disturbances. That is,

$$E[v_{it}|\mathbf{x}_i^t, \tilde{y}_i^{t-1}] = 0 \qquad (t = 2, ..., T) \tag{20}$$

$$E[\eta_i|\mathbf{x}_{it}] = E[\eta_i|\mathbf{x}_{is}] \qquad (t = 1, ..., T; s = 1, ..., T) \tag{21}$$

Constant correlation implies the moments

$$E[\mu_{it} \triangle \mathbf{x}_{it}] = 0 \tag{22}$$

in addition to those given by Assumption 2 (Predetermined Inputs). These new instruments, $\triangle \mathbf{x}_{it}$, augment the levels equation by providing instruments for cross-section inputs. In practice we must assume that any time-invariant inputs are uncorrelated with the fixed effect. Otherwise, the levels equation, which includes the time-invariant inputs, is not fully identified.

3.3.3 Levels GMM: Conditional mean stationarity

In some instances, particularly at later stages in a child's education, it may be reasonable to assume that while growth heterogeneity exists it does not speed rate of learning (rather common time effects account for unexplained achievement growth). This situation arises, for example, when the initial conditions have reached a convergent level with respect to the fixed effect such that,

$$y_{i1} = \frac{\eta_i}{1 - \beta} + d_i \tag{23}$$

where $t = 1$ is the first observed period not the first period in the learning life-cycle. Blundell and Bond (1998) discuss this type of conditional mean stationarity restriction in considerable depth. As they point out, the key assumption is that initial deviations d_i are uncorrelated with the level of $\eta_i/(1 - \beta)$. It does not imply that the achievement path $\{y_{i1}, y_{i2}, \ldots, y_{iT}\}$ is stationary; inputs, including time dummies, continue to spur achievement and can be nonstationary. The assumption only requires that conditional all the full set of controls and common time dummies, the fixed effect does not accelerate learning.[13]

[13]Note that even if this assumption fails, the moments it implies reduce the upward bias associated with omitted heterogeneity.

Assumption 5 (Conditional Mean Stationarity) *Learning growth is uncorrelated with the omitted heterogeneity. That is,*

$$E[\mu_{it} \triangle y_{i,t-1}] = 0 \qquad (t = 3, ..., T) \tag{24}$$

The conditional mean stationarity assumption provides an additional $T-2$ non-redundant moment conditions that can augment our system GMM estimators. While a fully efficient approach uses these additional moments along with typical moments in the differenced equation, the conditional mean stationarity assumption ensures strong instruments in the levels equation to identify β. Thus, if we prefer simplicity over efficiency, we can estimate the model using levels GMM or 2SLS and avoid the need to use a system estimator. Since educational researchers may be interested in simpler approaches, we instrument the undifferenced value-added model (5) using lagged changes in achievement $\triangle \tilde{y}_i^{t-1}$ and either changes in inputs $\triangle \mathbf{x}_i^t$ or inputs directly $\mathbf{x}_i^t$ depending on whether inputs are constantly correlated or uncorrelated with the individual effect.

4 Data

To demonstrate the issues we discuss, we use data collected by the authors as part of the Learning and Educational Achievement in Punjab Schools (LEAPS) project, an ongoing survey of learning in Pakistan. The sample comprises 112 villages in three districts of Punjab—Attock, Faisalabad and Rahim Yar Khan. The districts represent an accepted stratification of the province into North (Attock), Central (Faisalabad) and South (Rahim Yar Khan). Because the project was envisioned in part to study to dramatic rise of private schools in Pakistan, the 112 villages in these districts were chosen randomly from the list of all villages with an existing private school. As would be expected given the presence of a private school, the sample villages are generally larger, wealthier, and more educated than the average rural village. These type of villages, however, are spreading rapidly and represent the primary areas of change in the Pakistan context (Andrabi, Das and Khwaja, 2006).

The survey covers all schools within the sample village boundaries *and* within a short walk of any village household. Including schools that opened and closed over the three rounds, 858 schools were surveyed and only three refused to cooperate. Sample schools account for over 90% of enrollment in the sample villages.

The first panel of children consists of 13,735 third-graders, 12,110 of which were tested in Urdu,

English, and mathematics. These children were subsequently followed for two years and retested in each period. Every effort was made to track children across rounds, even when they were not promoted. In total, 12 percent and 13 percent of children dropped out or were lost between rounds one and two, and two and three, respectively. In addition to being tested, 6,379 children—up to ten in each school—were randomly administered a survey including anthropometrics (height and weight) and detailed family characteristics such parental education and wealth (as measured by PCA analysis of 20 assets).

For our analysis we use two basic subsamples of the data: all children who were tested and children who were tested *and* given a detailed child survey. Table 2 presents the characteristics of these children split by whether they attend public or private schools. The patterns across each subsample is relatively stable. Children attending privates schools are slightly younger, have fewer elder siblings, and come from wealthier and more educated households.

The measures of achievement are based on relatively lengthy exams in English, Urdu (the vernacular) and mathematics. Children were tested in third, fourth, and fifth grade during the Winter, at roughly one year intervals. Because the school year ends in the early spring, the test scores gains from third to fourth grade are largely attributable to the fourth grade school. The tests were scored by the authors using Item Response Theory so that the scale has cardinal meaning.[14] Preserving cardinality is important for longitudinal analysis since many other transformations, such as the percent correct score or percentile rank, are bounded artificially by the transformations that describe them. By comparison, IRT scores ensure that change in one part of the distribution is equal to a change in another, in terms of the latent trait captured by the test.

5 Results

5.1 Preliminary Results

5.1.1 Baseline estimates from cross-section data

Complementing the visual evidence of a large public-private school gap in Figures 1 and 2, Table 3 presents results for a cross-section regression of third grade achievement on child, household, and

[14] All items were modeled using the three parametric logistic (3PL) item response function and estimated using BILOG-MG.

school characteristics. These regressions provide some initial evidence that the public-private gap is more than omitted variables and selection. Adding a comprehensive set of child and family controls reduces the estimated coefficient on private schools only slightly. Adding village fixed effects also does not change the coefficient. Across all baseline specifications the gap remains large: over 0.9 standard deviations in English, 0.5 standard deviations in Urdu, and 0.4 standard deviations in math.

Besides the coefficient on school type, few controls are strongly associated with achievement. By far the largest other effect is for females, who outperform their male peers in all subjects except math. Height, assets, and whether the father (and for column 3, mother) is educated past elementary school also enter the regression positive and significant. More elder brothers correlates with lower achievement. These results confirm mild positive selection into private schools, but also suggest that, controlling for school type, few other observables seem to matter for achievement.

5.1.2 Graphical evidence from switching children

Many of the value-added estimators we explore identify the private school impact using children who switch schools. Figure 3 illustrates the patterns of achievement for these children. For each subject we plot two panels, the first containing children who start in public school and the second containing those who start in private school. We then graph achievement patterns for children who never switch, switch in after third grade, and switch after fourth grade. For simplicity, we exclude children who switch back and forth between school types.

As the table on the bottom of the figure shows, few children change schools. Only 48 children move to private schools in fourth grade and 40 move in fifth grade. Consistent with the role of private schools serving primarily younger children, 167 children switch to public schools in fourth grade and 160 switch in fifth grade. These numbers are roughly double the number of children available for our estimates that include controls, since only a random subset of children were surveyed regarding their family characteristics. The paucity of children switching schools underlies the primary difficulty comparing switching models—large standard errors.

Even given the small number of children switching school types, Figure 3 provides preliminary evidence that the private school effect is not simply a cross-section phenomenon. In all three subjects, children who switch to private schools between third and fourth grade experience large

achievement gains. Yet these gains are limited to the subsequent grade only. Children who switch between third and fourth grade experience rapid gains up till fourth grade as they converge to their new private school peers, but then learn at a rate similar to children in public schools. There are two competing interpretations: rapid gains may be due to a temporary shock that simultaneously speeds learning and induces switching, or once children converge to a higher achievement level the absolute level of depreciation increases making it more difficult to show large gains—consistent with $\beta < 1$ and Figures 1 and 2. One piece of evidence that supports the latter interpretation is the parallel trends for children who switch in fourth grade. There is no evidence that children experience any shock prior to shifting; achievement growth between third and fourth grade is the same for children who switch after fourth grade and children who never switch.

Children switching from private schools to public schools exhibit similar achievement patterns, except reversed. Moving to a public school is associated with slower learning or even learning losses. Unlike public schools, children switching out of private schools typically score below their private school peers. Again, most gains or losses occur immediately after moving; once achievement converges to the new level, children experience parallel growth in public and private schools. Again, this may be do either to a temporary shock or depreciation.

5.1.3 Addressing measurement error

Value-added learning models that include lagged achievement must confront the inherent measurement error associated with test scores. Table 4 applies the various approaches we discuss in Section 3.2 to our data. Each cell contains the estimated coefficient on lagged achievement from a regression with no controls and the associated standard error. Where applicable, we also report the p-value for Hansen's overidentification test statistic. This is possible for the instrumental variables estimators since we have three subject tests and three years of data.

Absent any correction (OLS), the estimated depreciation coefficient ranges between 0.65 and 0.70. Instrumenting using alternate subjects raises the estimated coefficient significantly to 0.85 for English, 0.89 for math, and 0.97 for Urdu. However, we can reject the overidentifying restriction at the 1 percent level in all three subjects. This suggests that measurement errors may be correlated across subjects at the same sitting and that this correlation may differ depending on the subject. By comparison, when we instrument for lagged achievement using double lagged scores we cannot reject

the overidentifying restrictions. Unfortunately, in the context of dynamic panel methods additional lags to address measurement error require $T = 4$. The final line of Table 5 shows estimates based on our analytical correction around 0.9. (Of course all of these estimates remain biased upward by growth heterogeneity.)

5.2 OLS and Dynamic Panel Value-Added Estimates

Tables 5 (English), 6 (Urdu), and 7 (math) summarize our main value-added results. All estimates include the full set of controls in the child survey sample, and the survey date, round (grade) dummies, and village fixed effects. For brevity, we only report the depreciation and private school coefficients. As discussed, time-invariant controls drop out of the differenced models. We also do not instrument using leads and lags of any structural controls such as age, years of schooling, survey date, and time and village dummies. For the system and levels estimators we also assume, by necessity, that time-invariant controls are uncorrelated with the fixed effect, or act as proxy variables.

Looking at the depreciation coefficient, we can immediately reject that achievement does not decay ($\beta = 1$). Across all specifications and subjects, the estimated depreciation coefficient is significantly lower than one, even in the specifications that correct for measurement error only and should be biased upward.

Assuming no depreciation biases the private school coefficient downward. For English, the estimated private school effect in the restricted model (first line) is negative and significant. For Urdu and math, we cannot reject that private schools have no effect on achievement. By comparison, all of our other estimates are positive and statistically significant, with the exception of the predetermined inputs difference GMM estimate which is too weak to identify the private school effect with any precision.

The typical lagged value-added model, which assumes no effects and measurement error, returns estimates between 0.52 and 0.58 for the deprecation coefficient. Correcting only for measurement error by instrument using the two alternate subjects—no effects (2SLS)—or by using the analytical correction described by in the appendix—no effects (HEIV)—increases the depreciation coefficient to between 0.70 and 0.79, *consistent with significant measurement error attenuation.* While the HEIV and 2SLS estimates agree, an overidentification test rejects exogeneity of the instruments

for English and math at the 5 and 10 percent levels, respectively. Notably, we cannot reject the overidentifying restrictions for math where both instruments are measures of verbal ability. As discussed earlier, this provides some evidence that the behavior of alternate subjects as instruments is dissimilar. Consistent with our analytical bias expressions and the results for the restricted value-added model, increasing the depreciation estimate reduces the private school effect—but there is no reason to think this is a better estimate.

Moving to our dynamic panel estimators, Panel B of each table gives the Arellano and Bond (1991) difference GMM estimates under the assumption that inputs are strictly exogenous or predetermined. These estimates correct for measurement error in the differenced model using the analytical correction described. Even so, instrumenting widens our confidence intervals substantially and makes it difficult to find statistically significant differences. In English and Urdu, the deprecation coefficient falls to between 0.34 and 0.50 but is not statistically significant different than the lagged-value-added model.

The estimates are, however, statistically significantly different than models that correct for measurement error only. *In other words, omitted growth heterogeneity exists and biases the static estimates upward*. For math, the estimated depreciation coefficient drops to 0.09 and 0.12, considerably below all the other estimates. While this rate of decay seems unrealistic, we present some external evidence in Section 6 that such rapid decay is plausible, at least for short run gains. In all cases, the there is no statistically significant change in the private school coefficient, although precision is limited.

For many of the GMM estimates, Hansen's J test rejects the overidentifying restrictions implied by the model. This rejection hints that either the model misspecified or some of the instruments are endogenous. This is troubling but not entirely unexpected. Many of the models we estimate have many overidentifying restrictions. One strategy to address this would be to slowly reduce the instrument set until the overidentification test is accepted or the model becomes just identified. We explored this strategy but found neither a clear story nor changes in our estimates. In Section 6.1 we do provide some evidence for strict exogeneity—households do not seem to react significantly to achievement shocks—since this is directly relevant to how one interprets the depreciation coefficient.

Moving to Panel C (levels and differences GMM) illustrates the benefit of a systems approach. In differences GMM we can only identify the private school effect with any precision if we assume

attendance is strictly exogenous. Yet this assumption will fail if shocks affects future enrollment decisions. Adding a levels equation using the assumption that inputs are constantly correlated or uncorrelated with the omitted effects, reduces that standard errors for the private school coefficient while maintaining the assumption that inputs are predetermined but not strictly exogenous. Under the reasonable scenario that inputs are constantly correlated with the omitted effect, the private school coefficient is large (0.32 to 0.38) and statistically significant. While we still cannot reject the hypothesis that it is equivalent to the lagged value-added model, the assumptions required for this estimator are far milder.

With the addition of a conditional mean stationarity assumption (Panel D), we can more precisely estimate the depreciation coefficient. In this model, we only use moments in levels to illustrate a dynamic panel estimator that improves over the lagged value-added model estimated by OLS but remains simple (it's simply a univariate 2SLS regression!). The depreciation coefficient rises slightly for Urdu and more significantly for math. This upward movement is consistent with a violation of the stationarity assumption (the fixed-effect still contributes to learning growth) but an overall reduction in the omitted heterogeneity bias. Again, the private school coefficient remains large and positive but statistically indistinguishable from the lagged value-added model estimated by OLS.

An overarching theme in the preceding analysis is that the assumed depreciation coefficient influences the estimated private school effect, but the precise exogeneity conditions matter little. In Figure 4, we graph the relationship between both coefficients explicitly. Rather than estimating the depreciation coefficient, we assume a specific rate and then estimate the value-added model. That is, we use $y_{it} - \beta y_{i,t-1}$ as the dependent variable. This has a number of advantages since it provides a robustness check for any estimated effects, requires only two years of data, and eliminates the need for complicated measurement error corrections (it assumes, however, that inputs are uncorrelated with the omitted growth heterogeneity). As expected, the estimated private school effect strongly depends on the assumed depreciation rate. Moving from the restricted value-added model ($\beta = 1$) to the pooled cross-section model ($\beta = 0$) increases estimated effect from negative or insignificant to large (over 0.5 standard deviations) and significant. Clearly our conclusions about the relative efficacy of private schools will depend critically on our beliefs about achievement dynamics.

5.3 Robustness Checks

Our estimates imply rapid achievement decay. Absent any educational inputs whatsoever, children lose over half of their achievement in a single year. For some subjects, such as math, this fraction may be even larger. While the estimates reported may strike some as implausibly high, they do fit a number of facts: the rapid fade-out observed in most educational interventions (discussed further in Section 6), the memory literature in psychology and neuroscience, and the known losses children experience over summer. All these anecdotes provide some evidence that for every two steps forward we might, as our estimates imply, take one steps back.

5.3.1 Expected bias with correlated omitted inputs

We can also get of sense for whether our estimates are reasonable by exploring the magnitude of the potential bias in a basic lagged value-added model. Consider, for example, the bias in the most basic regression $y_{it} = \alpha + \beta y_{i,t-1} + \eta_{it}$, where we have omitted all potential inputs and have already solved the problem of measurement error bias. Our estimates of this model suggest that the depreciation coefficient is at most 0.8 to 0.9—far higher than our dynamic panel estimates of around 0.5. Is this discrepancy reasonable?

Aggregating all the omitted contemporaneous inputs into one variable η_{it} implies the upward bias of the depreciation coefficient is simply $\text{Cov}(\eta_{it}, y_{i,t-1}) / \text{Var}(\dot{y_{it-1}})$. If the correlation between inputs in any two periods is a constant ρ_X, and all children in grade zero start from the same place, some simple algebra shows that, for example, the coefficient of depreciation in a lagged-value-added model will be biased upward to

$$\frac{\text{Cov}(\eta_{i4}, y_{i3})}{\text{Var}(\dot{y_{i3}})} = \frac{\rho_X}{2\beta\rho_X - \beta + \beta^2 + 1}. \tag{25}$$

Figure 5 gives a graphical representation of this bias calculation. To read the graph, choose a true depreciation coefficient β (the dotted lines) and a degree of correlation of inputs over time ρ_X (the horizontal axis). Given these choices, the y-axis reveals the depreciation coefficient a lagged value-added specification estimated by OLS would yield. Working with our estimates, we can see that if the true effect is 0.4 and inputs are correlated only 0.6 over time, the estimated effect will be fully 0.9. Given that the vast majority of inputs are fixed, this seems quite reasonable, and perhaps

even too low.

5.3.2 Bounding the depreciation coefficient using selection on observables

Another way to get at the reasonableness of rapid depreciation is motivated by Altonji, Elder and Taber's (2005) (hence AEL's) assumption of equal selection on observed and unobserved variables. AEL's novel contribution is the observation that under strict but not unreasonable assumptions, we can learn something about omitted variable bias by exploring how this bias would increase had we omitted the variables we do observe.

In contrast to AEL's application to Catholic schools, we are interested in the bias of a continuous variable and our dependent variable, achievement, is measured with error. Extending the model to account for these differences is trivial. Letting ε_{it} be the measurement error, an analogous assumption of equal selection on observables and unobservables is

$$\frac{\text{Cov}(\eta_{it}, y_{i,t-1})}{\text{Var}(\eta_{it}) - \text{Var}(\varepsilon_{it})} = \frac{\text{Cov}(\boldsymbol{\alpha}'\mathbf{x}_{it}, y_{i,t-1})}{\text{Var}(\boldsymbol{\alpha}'\mathbf{x}_i)}. \tag{26}$$

This condition states that the normalized shift in the observed components contribution to the dependent variable is equal to the normalized shift in unobserved components. The only difference to AEL's original application is we use the standardized coefficient rather than a standardized shift in the mean, and normalize the unexplained portion by the variation that is theoretically explainable.[15] Of course equal selection may not be true. If we intentionally select the covariates most likely to reduce omitted variable bias, the selection on unobservables would be less than the selection on observables. Nevertheless, this approach provides a useful heuristic to gauge the extent of bias due to unobservables.

Table 7 summarizes the results of this exercise. For each subject (columns) we run two lagged value-added regressions, one without controls and one with a full set of controls. Focusing on the first four rows and English, we can see the intuition behind AEL's suggestion. Absent controls the depreciation coefficient is 0.91 while the R^2 of the regression is 0.52. Adding controls raises the R^2 only modestly to 0.56 but at the same time reduces the estimated depreciation coefficient to

[15] Note that if $\eta_{it} = \eta^*_{it} + \varepsilon_{it}$ then the maximum unexplained variation that can be explained is $\text{Var}(\eta^*_{it})$, the non-measurement error variation. Since ε_i is measurement error, η^*_{it} and ε_{it} are uncorrelated and $\text{Var}(\eta^*_{it}) = \text{Var}(\eta_{it}) - \text{Var}(\varepsilon_{it})$.

0.74. Thus just by explaining an additional 4% of the total variation we reduced the depreciation coefficient substantially. This indicates that, as suspected, omitted inputs bias the depreciation coefficient upward.[16]

The last two lines of Table 7 show the bias corrected estimate that results from assuming equal selection on observed and unobserved variables. These results are nothing more than a formal extension of our ad-hoc comparison of how the R^2 and coefficient change when we drop controls. In all three subjects, the bias corrected estimate, which should be viewed more as a bound than an actual estimate, is less than 0.05. Our dynamic panel estimates therefore appear reasonable; they are consistent with the degree of bias we encounter when omitting the variables we do observe.

6 Interpretation and Implications

6.1 Interpretation: Concavity, Smoothing or Forgetting?

The estimation issues discussed above are largely unrelated to the economic interpretation of the lagged test-score estimate. We have used the term depreciation to refer to the estimate on the lagged test-score in a linear panel data model simply because it is consistent with our learning framework. But this term jumps ahead of the analysis so far presented. In a simple optimizing model of behavior, the lagged coefficient estimate can be interpreted as "depreciation" only if (a) preferences over test-scores are linear and (b) the educational production function is separable in the stock of knowledge and future learning. Violations of (a) imply that the lagged-test score coefficient could arise from household re-optimization following a positive shock or heterogeneity in initial endowments. Violations of (b) imply that the coefficient could arise from concavity in the educational production function—when you know more, it costs more to learn. Understanding the channels through which this coefficient is obtained is critical. If the coefficient arises from pure depreciation, policy innovations such as summer school or shorter breaks through the year may have a large effect. If the coefficient arises from household optimization behavior or underlying concavities, there may be no option but to live with it.

Smoothing of achievement shocks may stem from household or school reactions. At the household

[16] We use the empirical Bayes estimate of lagged achievement that incorporates information contained in the control variable set to correct for measurement error. This ensures that adding controls does not simply capture some of the effect of lagged achievement on future learning, because achievement is noisily measured.

level, parents may substitute away inputs from children who are above their target achievement level.[17] At the school level, teachers may target struggling children for special attention. Tables 8 and 9 explore these two possibilities. To test household behavior responses, we examine whether inputs adjust to unexpected achievement shocks for roughly 650 children for whom we have detailed information from a survey collected at households. As a measure of the unexpected shock, we first compute the residual from a regression of fourth grade scores on third grade scores and a host of known controls. We then test whether this residual predicts changes between fourth and fifth grade in parents' perceptions of the child's performance and five educational inputs.[18] Table 8 presents the results. While parents' perceptions of their child's performance reacts strongly to gains to achievement, we find limited, if any, input changes. School expenditure drop slightly as do the hours spent helping the child on his or her homework; minutes spent playing increases. Few responses, however, are statistically significant.

Table 9 explores the possibility that the depreciation captures teachers targeting poorly performing students. Here, we estimate the basic lagged value-added model with no controls, and instrument for lagged achievement using lagged differences in alternate subjects (i.e. the basic moments from conditional mean stationarity). To see whether concavity is a within or between school phenomenon, we estimate the model using mean school scores (between) and school-demeaned child scores (within). If teachers target poorly performing children in each classroom, we'd expect the depreciation coefficient to be lower within schools than between schools. In fact, we find the opposite. If anything, the depreciation coefficient is lower for the between school regressions.

A final possibility is that children simply forget what they learned, consistent with our "depreciation" terminology. Perhaps the most convincing evidence in this regard comes from psychology and neuroscience, the literature on summer learning loss, and experimental evidence of program fade out. In psychology, research on the "curve of forgetting" dates back Ebbinghaus's (1885) seminal study on memorization and forgetting of nonsense syllables. Rubin and Wenzel (1996) review the massive outpouring of laboratory research spawned by this contribution. Semb and Ellis (1994) review classroom studies—how much students remember after taking a course. While much of this

[17] Alternatively, if there is strong complementarity between inputs and achievement (e.g. Cunha and Heckman, 2007*b*) parents may increase inputs when children perform well. At least in the short run, our results suggests this effect does not dominate.

[18] We instrument for the subject specific residuals using the alternate subject residuals to lessen measurement error attenuation.

research explores the form and nature of forgetting, there is one common theme: learning quickly fades absent continued study.

In a similar vein, educationalists have long been concerned by summer learning loss. Even with the stimulus many summer activities supply, children typically experience learning losses between spring and fall achievement tests (Cooper et al., 1996). These losses are generally not as rapid as the effects we find, but the experiment is different: we estimate the depreciation with no inputs whereas summer activities provide some stimulus, particularly for privileged children.

In addition to laboratory experiments, field experiments in education provide an ideal test of achievement decay, although it is often difficult to rule out competing explanations. Table 10 summarizes six randomized (or quasi-randomized) interventions that followed children after the program ended. This follow-up enables estimation of both immediate and extended treatment effects. For the interventions summarized, the extended treatment effect represents test scores roughly one year after the particular program ended.

Banerjee et al. (2007) report on two interventions in India. The Balsakhi program provided young and relatively inexperienced "balsakhis" (tutors) for children who struggled in school. In terms of the immediate treatment effect, the program was a considerable success. Average achievement in treatment schools, including children who did not directly receive help, leapt by 0.34 standard deviations in math and 0.23 standard deviations in verbal. A year after the program ended, however, virtually no effect remained. The implied depreciation coefficient is below 0.1 and far below our estimates, even for math. A second "computer assisted learning" program (CAL) experienced a similar pattern of high initial treatment effects but then considerable decay ($\beta = 0.265$).

Glewwe, Ilias and Kremer (2003) and Kremer, Miguel and Thornton (2003) report outcomes for an incentive program for teachers and students, also within the developing country context. Both programs saw immediate impacts above 0.1 standard deviations. After a year, however, the teacher incentive program gains became statistically insignificant and the student incentive gains dropped 30 percent. While these estimates support rapid depreciation, they are contaminated somewhat by dynamic treatment effects; Kremer, Miguel and Thornton (2003) argue that the immediate effects of teacher incentives were primarily driven by manipulation and Glewwe, Ilias and Kremer (2003) argue that student incentives increased effort (not just achievement) even after the program ended.

In the United States, Krueger and Whitmore (2001) review evidence from the STAR class size

experiment and conclude achievement gains one year after the program ended fell to between a quarter and a half of their original levels. Moreover, these estimates are optimistic because children who experienced the treatment typically enrolled in smaller classes even after the program ended. Jacob and Lefgren (2004) estimate the immediate and extended treatment effect of a summer school and grade retention policy in Chicago public schools using a quasi-experimental regression-discontinuity design. In line with rapid fade-out, they find a two year impact 30 to 40 percent lower than the immediate impact following summer school and retention. If retention benefits children even after the first year, as retention advocates claim, the implied depreciation rate would rise further.

6.2 Implications for Experimental and Non-Experimental Program Evaluation

In the absence of randomized studies, the value-added approach has gained momentum as a valid methodology for removing unobserved individual heterogeneity in assessing the contribution of specific programs, or understanding the contribution of school-level factors for learning. Our results reject both the assumption of zero depreciation required for the restricted value-added model and no growth effects required for the lagged value-added model. Individual heterogeneity affects learning over time. In other words the original goal of value-added models—to eliminate individual effects—remains unaccomplished.

For non-experimental program evaluation the conclusion here is mixed. Perhaps our most striking result is the remarkable failure of the restricted value-added model; the time has clearly come to abandon it forever. In our application, cross-sectional analysis is more reliable than the fundamentally misspecified gain-score model. Perhaps one reason for its continued application is the well-known difficulties associated with lagged dependent variables. As we have shown, the dynamic panel literature is filled with fruitful suggestions and solutions. Some of these estimators are easy to implement and researchers estimating value-added models would be well served in doing so. But they are not a panacea. All the estimators we present require either exogenous (at least in the predetermined sense) time-series input variation or uncorrelated effects.

More pessimistically, our results also highlight difficulties with randomized evaluations. The extent of depreciation implies that a short evaluation yields little information about the cost-effectiveness of a program. Using the one or two year increase from a program only gives an

upper-bound on the longer term gains. As our estimates suggest, and Table 10 confirms, we should expect program impacts to fade quickly. Calculating the internal rate of return by citing research linking test scores to earnings of young adults is therefore a doubtful proposition. The techniques described here, with 3 periods of data (a luxury in many evaluations), can theoretically obtain a lower-bound on cost-effectiveness by assuming exponential depreciation. But doing so seems to be taking our model too literally and generally yields uninformative bounds anyway.

Consider the popular strategy of summarizing results in tenths of test score standard deviations gained per dollar allocated. Such comparisons are ultimately undermined if depreciation varies across interventions as Table 10 suggests it might.[19] In the education literature, there is also considerable evidence for heterogeneous decay across treatments (Semb and Ellis, 1994). To give one example, MacKenzie and White (1982) report on an experiment where children learned geographical knowledge either during an in-class exercise or field excursion. They find that while initial gains were roughly equivalent, after 12 weeks the first group lost over 40 percent of their knowledge compared to 10 percent for the active learning group. This lesson doesn't just apply to comparing educational methods. Rothstein (2007) finds heterogeneity in the long-run effects of teachers who produce equal short-run gains—raising significant doubts about value-added accountability systems.

Using our data, we briefly examine depreciation heterogeneity in Table 11. Here, we estimate the value-added model for specific sub-populations using the "predetermined inputs, uncorrelated effects, and conditionally stationary" based estimator (last line of Table 5, 6, and 7). Unfortunately, large standard errors make it difficult to find statistically different decay rates between groups. Learning in private schools seems to decay faster than learning in public schools, but the difference is not statistically significant. A similar pattern holds for richer families and children with educated parents. While more research is required, these results hint that learning decays faster for faster learners.

As another strategy, many researchers compare interventions to the average "yearly gain" (e.g. Angrist et al., 2002; Neal, 2002; Jacob and Lefgren, 2004; Banerjee et al., 2007); a program is equal

[19]There are other problems with comparing interventions using standard deviations as well. Most randomized evaluations use ad-hoc tests that are not comparable. While standardizing removes some of the bias, doing so does not change the form of the score distribution. It many cases it may therefore be possible to construct exams that yield different results. Perhaps more importantly, each normalization divides by a standard deviation that is not comparable; a village in India does not have the same score variance as all of Punjab and a short test will produce scores with more variance than a long test simply due to measurement error.

to "5 months of school," say. But when achievement decays, observed gains do not equal total gains. A child who scores 100 in third grade and 120 in fourth grade has learned 20 *plus* whatever she forgot. In fact, the amount forgotten is unidentifiable in an absolute sense because "zero" knowledge is undefined. Using the minimum score as a bound and assuming proportional decay, the true gain is likely several times the observed gain. While the observed gain over a year is sometimes a useful concept, expressing treatment effects as a percentage of the "average yearly gain" tends to exaggerate the impact compared to other inputs.

7 Conclusion

The conclusion that input based policies contribute little to achievement is driven in part by results from value-added specifications. These specifications are considered superior to cross-sectional comparisons because they presumably difference out the omitted effect of fixed household and child characteristics. As we have shown, this view is incorrect. The restricted value-added specification assumes that lagged achievement carries over with no loss. Our data strongly rejects this assumption. This rejection is not new. All the lagged value-added specification in the literature we reviewed report remarkably similarly estimates for the depreciation coefficient yet gain-score models remain common. Our results for Pakistan should illustrate the danger: the restricted value-added model is fundamentally misspecified and performs worse than cross-sectional analysis.

Beyond the dramatic rejection of the restricted model, our main contribution is to highlight the estimation problems associated with the lagged value-added specification and to survey dynamic panel estimators that overcome the primary difficulties involved. As typically estimated, the depreciation coefficient is biased downward by measurement error and upward by omitted individual heterogeneity. The degree to which these biases cancel depends on the particular dataset. While we find the lagged value-added model yields reasonable estimates, we strongly reject the assumption that individual heterogeneity does not speed learning.

The rapid rate of depreciation we find—around 0.5 for English and Urdu and potentially lower for mathematics—is consistent with analytical and empirical estimates of the expected bias under OLS and with experimental evidence of program fade-out in developing and developed countries. We also present evidence that this decay is not solely due to smoothing of achievement shocks by

parents and teachers. Achievement, at least in so far as measured by test scores, fades rapidly barring continued study.

The implications of rapid depreciation are broader than choosing an appropriate estimator for value-added models. Many techniques in program evaluation are no longer strictly legitimate. Any effort to calculate internal rates of return using short-run achievement data is generally hopeless. At least in terms of achievement, one time gains are quickly erased. Those effects that do remain are more likely due to dynamic treatment effects—e.g. self-productivity of early childhood education (Heckman and Masterov, 2007)—than the immediate impact of a particular program on achievement. Short-run gains are poor proxies for ultimate impacts. Determining which programs have such dynamic and long-run effects is an exciting area for future research.

More generally, rapid depreciation suggests learning is not a linear accumulation of knowledge. Indeed, forgetting may be a natural part of the learning process. As the great British philosopher Alfred Whitehead opined "A merely well informed man is the most useless bore on God's earth." Even if this statement is too strong, recent work by Todd and Wolpin (2006) and Cunha and Heckman (2007*b*) indicates that the separable value-added model is inconsistent with the actual learning process. Likewise, research by Bowles, Gintis and Osborne (2001) and Heckman and Rubinstein (2001), among others, suggests that achievement in specific domains captured by formal exams may be of secondary importance to non-cognitive skills. Combined with our results, this research pleads for a richer model of education. Not simply to add nuance to our understanding of learning, but to get the most basic parameters right.

A Analytical Corrections for Measurement Error

Consider the lagged value-added model

$$y_{it} = \alpha' \mathbf{x}_{it} + \beta y^*_{i,t-1} + v_{it}, \tag{27}$$

where $y^*_{i,t-1}$ is the true lagged achievement, v_{it} is the error term, and we have put aside the possibility of omitted heterogeneity. Since $y^*_{i,t-1}$ is a latent variable, we can only estimate it with error. Thus we actually estimate

$$y_{it} = \alpha' \mathbf{x}_{it} + \beta y_{i,t-1} + [v_{it} - \beta \varepsilon_{i,t-1}] \tag{28}$$

and OLS is inconsistent.

The analytic correction we apply, replaces $y_{i,t-1}$ with the best linear predictor

$$\tilde{y}_{i,t-1} \equiv E[y^*_{i,t-1} \mid y_{i,t-1}, \mathbf{x}_{it}] = \lambda' \mathbf{x}_{it} + r_{i,t-1} y_{i,t-1}, \tag{29}$$

where λ and $r_{i,t-1}$ are parameters. To see why this works, add and subtract $\beta \tilde{y}_{i,t-1}$ from (**??**) to get

$$y_{it} = \mathbf{x}_{it}\alpha + \beta \tilde{y}_{i,t-1} + [\beta(y^*_{i,t-1} - \tilde{y}_{i,t-1}) + v_{it}]. \tag{30}$$

OLS is consistent if

$$E[\mathbf{x}'_{it}(y^*_{i,t-1} - \tilde{y}_{i,t-1})] = 0, \tag{31}$$

$$E[\tilde{y}_{i,t-1}(y^*_{i,t-1} - \tilde{y}_{i,t-1})] = 0. \tag{32}$$

These conditions are automatically satisfied since the fitted value $\tilde{y}_{i,t-1}$ and independent variables $\mathbf{x}_{it}$ are orthogonal to the residual $y^*_{i,t-1} - \tilde{y}_{i,t-1}$ by the definition of a projection.

The only difficulty is estimating the projection parameters λ and $r_{i,t-1}$ since the dependent variable $y^*_{i,t-1}$ is unobserved. But it turns out that we do not need to observe the true score. The

orthogonality conditions that define the projection (29) are

$$E[\mathbf{x}'_{it}(y^*_{i,t-1} - \lambda'\mathbf{x}_{it} - r_{i,t-1}y_{i,t-1})] = 0, \tag{33}$$

$$E[y'_{i,t-1}(y^*_{i,t-1} - \lambda'\mathbf{x}_{it} - r_{i,t-1}y_{i,t-1})] = 0. \tag{34}$$

Solving first for λ, we have

$$\lambda = E[\mathbf{x}'_{it}\mathbf{x}_{it}]^{-1}E[\mathbf{x}'_{it}(y^*_{i,t-1} - r_{i,t-1}y_{i,t-1})]. \tag{35}$$

Plugging (35) into (34) and solving for $r_{i,t-1}$ yields

$$r_{i,t-1} = E[y_{i,t-1}\mathbf{m}_x y_{i,t-1}]^{-1}E[y_{i,t-1}\mathbf{m}_x y^*_{i,t-1}] \tag{36}$$

$$= E[e^2_{i,t-1}]^{-1}\left(E[e^2_{i,t-1}] - E[\varepsilon^2_{i,t-1}]\right) \tag{37}$$

$$= \frac{\sigma^2_{e_{i,t-1}} - \sigma^2_{\varepsilon_{i,t-1}}}{\sigma^2_{e_{i,t-1}}}, \tag{38}$$

where $\mathbf{m}_x \equiv 1 - \mathbf{x}_{it}(\mathbf{x}'_{it}\mathbf{x}_{it})^{-1}\mathbf{x}'_{it}$ is an annihilator vector and $e_{i,t-1}$ is the residual from a regression of $y_{i,t-1}$ on $\mathbf{x}_{it}$. We can estimate $r_{i,t-1}$ by computing $\sigma^2_{e_{i,t-1}}$ from the regression of $y_{i,t-1}$ on $\mathbf{x}_{it}$ and taking $\sigma^2_{\varepsilon_{i,t-1}}$ from IRT. Intuitively, $r_{i,t-1}$ is the heteroscedastic reliability ratio of the score minus the variation explained by the independent variables. That is, the reliability ratio of $y_{i,t-1} - E[y^*_{i,t-1} \mid \mathbf{x}_{it}]$.

We compute λ by plugging $r_{i,t-1}$ into (35) to get

$$\lambda = E[\mathbf{x}'_{it}\mathbf{x}_{it}]^{-1}E[\mathbf{x}'_{it}(y^*_{i,t-1} - r_{i,t-1}y_{i,t-1})] \tag{39}$$

$$= E[\mathbf{x}'_{it}\mathbf{x}_{it}]^{-1}E[\mathbf{x}'_{it}y_{i,t-1}](1 - r_{i,t-1}). \tag{40}$$

The best predictor is,

$$\tilde{y}_{i,t-1} = E[y_{i,t-1}|\mathbf{x}_{it}](1 - r_{i,t-1}) + r_{i,t-1}y_{i,t-1} \tag{41}$$

This takes the familiar form of an empirical Bayes estimate that shrinks the observed score to the predicted mean. The shrinkage performs the same function has blowing up the coefficient using the reliability ratio after estimation. Here, however, our shrunken estimate provides a more efficient

correction by using the full heteroscedastic error structure (Sullivan, 2001) .

In practice, our empirical Bayes estimate incorporates as much information as possible, including forward and backward input lags and forward and backward lagged test scores, in all subjects. While regressors that are not in the ultimate estimating equation need not be included to correct for measurement error, additional information provides more precise estimates, and allows us to use one empirical Bayes estimate across all specifications. The only requirement is that no regressor in the estimating equation is left out of the empirical Bayes estimate.

References

Altonji, J.G., T.E. Elder and C.R. Taber. 2005. "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools." *Journal of Political Economy* 113(1):151–184.

Anderson, TW and C. Hsiao. 1981. "Estimation of Dynamic Models with Error Components." *Journal of the American Statistical Association* 76(375):598–606.

Andrabi, Tahir, Jishnu Das and Asim Ijaz Khwaja. 2006. "A dime a day : the possibilities and limits of private schooling in Pakistan." *World Bank Policy Research Working Paper 4066* .

Angrist, J., E. Bettinger, E. Bloom, E. King and M. Kremer. 2002. "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment." *The American Economic Review* 92(5):1535–1558.

Arellano, M. 2003. *Panel Data Econometrics*. Oxford University Press.

Arellano, M. and O. Bover. 1995. "Another look at the instrumental variable estimation of error-components models." *Journal of Econometrics* 68(1):29–51.

Arellano, M. and S. Bond. 1991. "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations." *The Review of Economic Studies* 58(2):277–297.

Arellano, Manuel and Bo Honore. 2001. Panel data models: some recent developments. In *Handbook of Econometrics*, ed. J.J. Heckman and E.E. Leamer. Vol. 5 of *Handbook of Econometrics* Elsevier chapter 53, pp. 3229–3296.

Banerjee, Abhijit, Shawn Cole, Esther Duflo and Leigh Linden. 2007. "Remedying Education: Evidence from Two Randomized Experiments in India." *Quarterly Journal of Economics* 122(3).

Ben-Porath, Y. 1967. "The Production of Human Capital and the Life Cycle of Earnings." *The Journal of Political Economy* 75(4):352–365.

Bhargava, A. and JD Sargan. 1983. "Estimating Dynamic Random Effects Models from Panel Data Covering Short Time Periods." *Econometrica* 51(6):1635–1660.

Birnbaum, Allan. 1968. Some Latent Trait Models and Their Use in Inferring an Examinee's Ability. In *Statistical Theories of Mental Test Scores*, ed. Frederic M. Lord and Melvin R. Novick. Addison-Wesley Publishing Company.

Blundell, R. and S. Bond. 1998. "Initial conditions and Moment Conditions in Dynamic Panel Data Models." *Journal of Econometrics* 87(1):115–43.

Boardman, A.E. and R.J. Murnane. 1979. "Using Panel Data to Improve Estimates of the Determinants of Educational Achievement." *Sociology of Education* 52(2):113–121.

Bowles, S., H. Gintis and M. Osborne. 2001. "The Determinants of Earnings: A Behavioral Approach." *Journal of Economic Literature* 39(4):1137–1176.

Breusch, T.S., G.E. Mizon and P. Schmidt. 1989. "Efficient Estimation Using Panel Data." *Econometrica* 57(3):695–700.

Chay, K.Y., P.J. McEwan and M. Urquiola. 2005. "The Central Role of Noise in Evaluating Interventions That Use Test Scores to Rank Schools." *The American Economic Review* 95(4):1237–1258.

Cooper, H., B. Nye, K. Charlton, J. Lindsay and S. Greathouse. 1996. "The Effects of Summer Vacation on Achievement Test Scores: A Narrative and Meta-Analytic Review." *Review of Educational Research* 66(3):227–68.

Cunha, F. and J.J. Heckman. 2007*a*. "Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Journal of Human Resources* .

Cunha, F, JJ Heckman and SM Schennach. 2006. "Estimating the Elasticity of Substitution Between Early and Late Investments in the Technology of Cognitive and Noncognitive Skill Formation." *Unpublished, University of Chicago, Department of Economics* .

Cunha, Flavio and James Heckman. 2007*b*. "The Technology of Skill Formation." *American Economic Review* 97(2):31–47.

Currie, J. and D. Thomas. 1995. "Does Head Start Make a Difference?" *The American Economic Review* 85(3):341–364.

Deming, David. 2007. "Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start." Harvard University. Processed.

Doran, H. and L.T. Izumi. 2004. "Putting Education to the Test: A Value-Added Model for California." *San Francisco: Pacific Research Institute* .

Ebbinghaus, H. 1885. *Memory: A contribution to experimental psychology.* New York: Teachers College, Columbia University.

Garces, E., D. Thomas and J. Currie. 2002. "Longer-Term Effects of Head Start." *The American Economic Review* 92(4):999–1012.

Glewwe, P., N. Ilias and M. Kremer. 2003. "Teacher Incentives." *NBER Working Paper* .

Gordon, Robert, Thomas J. Kane and Douglas O. Staiger. 2006. "Identifying Effective Teachers Using Performance on the Job." *Hamilton Project Discussion Paper* .

Gronberg, T.J., D.W. Jansen and G.S. Naufal. 2006. Efficiency And Performance In Texas Public Schools. In *Improving School Accountability: Check-Ups or Choice*, ed. Timothy J. Gronberg and Dennis W. Jansen. Vol. 14 of *Advances in Applied Microeconomics* Elsevier.

Hanushek, E.A. 1979. "Conceptual and Empirical Issues in the Estimation of Educational Production Functions." *The Journal of Human Resources* 14(3):351–388.

Hanushek, E.A. 2003. "The Failure of Input-Based Schooling Policies." *Economic Journal* 113(485):64–98.

Hausman, J.A. and W.E. Taylor. 1981. "Panel Data and Unobservable Individual Effects." *Econometrica* 49(6):1377–1398.

Heckman, James J. and Dimitriy V. Masterov. 2007. "The Productivity Argument for Investing in Young Children." *NBER Working Paper 13016* .

Heckman, J.J. and Y. Rubinstein. 2001. "The Importance of Noncognitive Skills: Lessons from the GED Testing Program." *The American Economic Review* 91(2):145–149.

Imbens, Guido W. and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62:467–475.

Jacob, B. A. and L. Lefgren. 2004. "Remedial education and student achievement: A regression-discontinuity analysis." *Review of Economics and Statistics* 86(1):226–244.

Jacob, B. A. and L. Lefgren. 2005. "What Do Parents Value in Education: An Empirical Investigation of Parents' Revealed Preferences for Teachers." *NBER Working Paper 11494* .

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

Kremer, M., E. Miguel and R. Thornton. 2003. "Incentives to Learn." *NBER Working Paper* .

Krueger, A.B. 2003. "Economic Considerations and Class Size." *Economic Journal* .

Krueger, A.B. and D.M. Whitmore. 2001. "The Effect of Attending a Small Class in the Early Grades on College-test Taking and Middle School Test Results: Evidence from Project Star." *The Economic Journal* 111(468):1–28.

Ladd, H.F. and R.P. Walsh. 2002. "Implementing value-added measures of school effectiveness: getting the incentives right." *Economics of Education Review* 21(1):1–17.

MacKenzie, A.A. and R.T. White. 1982. "Fieldwork in Geography and Long-Term Memory Structures." *American Educational Research Journal* 19(4):623–632.

McCaffrey, D.F. 2004. *Evaluating Value-added Models for Teacher Accountability*. Rand Corporation.

Murnane, R.J., J.B. Willett and F. Levy. 1995. "The Growing Importance of Cognitive Skills in Wage Determination." *The Review of Economics and Statistics* 77(2):251–266.

Neal, D. 2002. "How Vouchers Could Change the Market for Education." *The Journal of Economic Perspectives* 16(4):25–44.

Neal, D. and W. Johnson. 1996. "The Role of Premarket Factors in Black-White Wage Differentials." *Journal of Political Economy* 104(5):869–895.

Rothstein, Jesse. 2007. "Do Value-Added Models Add Value? Tracking, Fixed Effects, and Causal Inference." *Working Paper* .

Rubin, DC and AE Wenzel. 1996. "One Hundred Years Of Forgetting: A Quantitative Description Of Retention." *Psychological Review* 103(4):734–760.

Santibanez, Lucrecia. 2006. "Why we should care if teachers get A's: Teacher test scores and student achievement in Mexico." *Economics Of Education Review* 25(5):510–520.

Sass, T.R. 2006. "Charter Schools and Student Achievement in Florida." *Education Finance and Policy* 1(1):91–122.

Schwartz, A.E. and J. Zabel. 2005. The Good, the Bad, and the Ugly: Measuring School Efficiency Using School Production Functions. In *Measuring School Performance and Efficiency: Implications for Practice and Research*, ed. L. Stiefel, A. E. Schwartz, R. Rubenstein and J. Zabel. NY: Eye on Education, Inc. pp. 37–66.

Semb, G.B. and J.A. Ellis. 1994. "Knowledge taught in school: What is remembered." *Review of Educational Research* 64(2):253–286.

Su, Xuejuan. 2004. "The Human Capital Dynamic Linkage in Early Childhood Development: How Pre-Kindergarten Experience Affects Schooling Outcomes." *Unpublished Manuscript: Economics, Finance and Legal Studies, The University of Alabama* .

Sullivan, D.G. 2001. "A Note on the Estimation of Linear Regression Models with Heteroskedastic Measurement Errors." *Federal Reserve Bank of Chicago* .

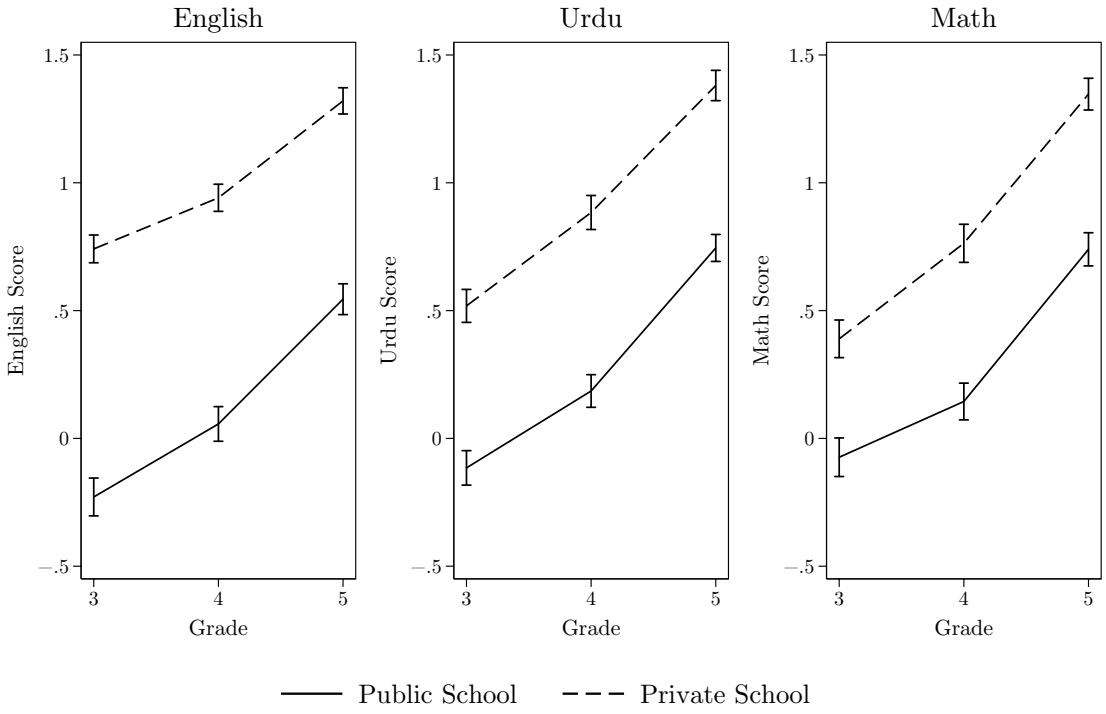
Todd, P.E. and K.I. Wolpin. 2003. "On the Specification and Estimation of the Production Function for Cognitive Achievement." *Economic Journal* 113(485):3–33.

Todd, P.E. and K.I. Wolpin. 2006. "The Production of Cognitive Achievement in Children: Home, School and Racial Test Score Gaps." *Philadelphia, PA: University of Pennsylvania, PIER Working Paper* pp. 04–019.

Whittemore, A.S. 1989. "Errors-in-Variables Regression Using Stein Estimates." *The American Statistician* 43(4):226–228.

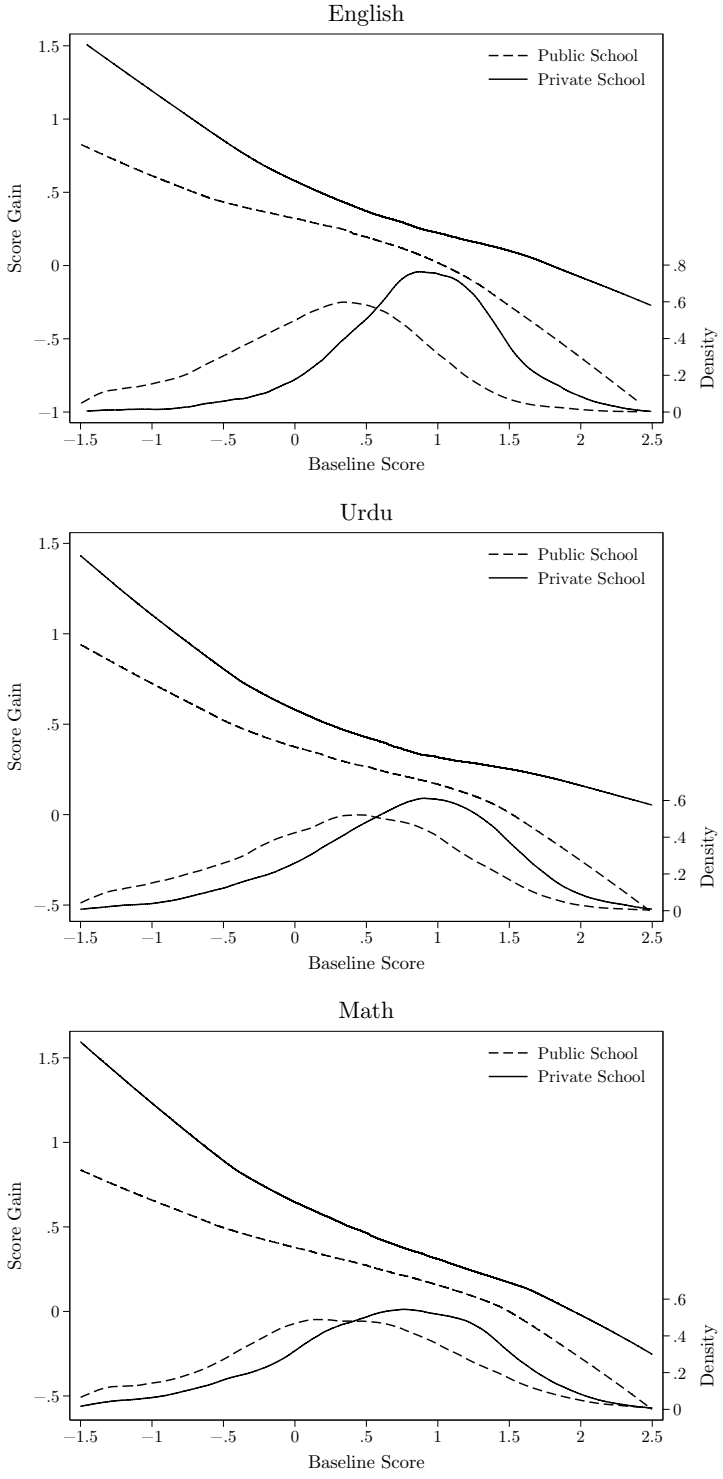
Wixted, J.T. and E.B. Ebbesen. 1991. "On the form of forgetting." *Psychological Science* 2(6):409–415.

FIGURE 1. EVOLUTION OF TEST SCORES IN PUBLIC AND PRIVATE SCHOOLS



Notes: Vertical bars represent 95% confidence intervals around the group means, allowing for arbitrary clustering within schools. Children who were tested in third grade were subsequently followed and counted as being in fourth or fifth grade regardless of whether they were actually promoted. The graph's sample is limited to children who were tested in all three periods (Table 2, Panel A: Full Sample).

FIGURE 2. ACHIEVEMENT GAINS AND BASELINE ACHIEVEMENT



Notes: The yearly score gain vs. baseline score lines for public and private schools are smoothed using a locally-weighted regression estimator. The sample includes children tested in all periods with baseline score between -1.5 and 2.5 (to avoid noise near tails). The kernel density for children's scores in public and private schools uses the scale on the right y-axis.

TABLE 1. DYNAMIC PANEL ESTIMATOR SUMMARY

Estimator Assumptions	"Difference" Instruments	"Levels" Instruments	Notes
Panel A: Static Estimates			
No depreciation β=1 (OLS)	n/a	n/a	Assumes no depreciation and no or uncorrelated heterogeneity.
No effects, no measurement error (OLS)	n/a	n/a	Assumes no measurement error and no effects.
No effects (2SLS)	n/a	Alternate subjects	Assumes no effects and uncorrelated measurement errors across subjects
No effects (HEIV)	n/a	n/a	Assumes no effects and analytical correction is valid.
Panel B: Difference GMM			
Strictly exogenous inputs	Inputs: 1...T Score: 1...t-2	n/a	Assumes no feedback effects.
Predetermined inputs	Inputs: 1...t-1 Score: 1...t-2	n/a	None (beyond those that apply to all estimators)
Panel C: Levels and Difference SGMM			
Predetermined inputs, constantly correlated effects	Inputs: 1...t-1 Score: 1...t-2	ΔInputs: 1..t	Assumes effects have constant correlation with inputs.
Predetermined inputs, uncorrelated effects	Inputs: 1...t-1 Score: 1...t-2	Inputs: 1..t	Assumes effects are uncorrelated with inputs (random effects).
Panel D: Levels GMM (Proxy Style)			
Predetermined inputs, constantly correlated effects, weakly stationary	Inputs: 1...t-1 Score: 1...t-2 (*not used)*	ΔInputs: 1..t ΔScore: t-1	Assumes effects have constant correlation with inputs and scores are conditionally mean stationary.
Predetermined inputs, uncorrelated effects, weakly stationary	Inputs: 1...t-1 Score: 1...t-2 (*not used*)	Inputs: 1..t ΔScore: t-1	Assumes effects are uncorrelated with inputs (random effects) and scores are conditionally mean stationary.

Notes: The notes columns do not include knife-edge cases such as perfectly offsetting biases. None of the dynamic panel estimators allow for serial correlation, as written. Redundant instruments in levels and differences are dropped. Panel D lists the valid difference instruments but our application does not use them in order to demonstrate a simple single equation estimator.

Table 2. Baseline Characteristics of Children in Public and Private Schools

Variable	Private School	Public School	Difference
Panel A: Full Sample			
Age	9.58	9.63	-0.04
	[1.49]	[1.35]	(0.08)
Female	0.45	0.47	-0.02
			(0.03)
English score (third grade)	0.74	-0.23	0.97***
	[0.61]	[0.94]	(0.05)
Urdu score (third grade)	0.52	-0.12	0.63***
	[0.78]	[0.98]	(0.05)
Math score (third grade)	0.39	-0.07	0.46***
	[0.81]	[1.00]	(0.05)
N	2337	5783	
Panel B: Surveyed Child Sample			
Age	9.63	9.72	-0.09
	[1.49]	[1.34]	(0.08)
Female	0.47	0.48	-0.02
			(0.03)
Years of schooling	3.39	3.75	-0.35***
	[1.57]	[1.10]	(0.08)
Weight z-score (normalized to U.S.)	-0.75	-0.64	-0.10
	[4.21]	[1.71]	(0.13)
Height z-score (normalized to U.S.)	-0.42	-0.22	-0.20
	[3.32]	[2.39]	(0.13)
Number of elder brothers	0.98	1.34	-0.36***
	[1.23]	[1.36]	(0.05)
Number of elder sisters	1.08	1.27	-0.19***
	[1.27]	[1.30]	(0.05)
Father lives at home	0.88	0.91	-0.04***
			(0.01)
Mother lives at home	0.98	0.98	0.00
			(0.01)
Father educated past elementary	0.64	0.46	0.18***
			(0.02)
Mother educated past elementary	0.36	0.18	0.18***
			(0.02)
Asset index (PCA)	0.78	-0.30	1.08***
	[1.50]	[1.68]	(0.07)
English score (third grade)	0.74	-0.24	0.99***
	[0.62]	[0.95]	(0.05)

Urdu score (third grade)	0.53	-0.14	0.67***
	[0.78]	[0.98]	(0.05)
Math score (third grade)	0.42	-0.09	0.51***
	[0.80]	[1.02]	(0.05)
N	1374	2657	

* Significant at the 10%; ** significant at the 5%; *** significant at 1%.

Notes: Cells contain means, brackets contain standard deviations, and parentheses contain standard errors. Standard errors for the private-public difference are clustered at the school level. Sample includes only those children tested (A) and surveyed (B) in all three years.

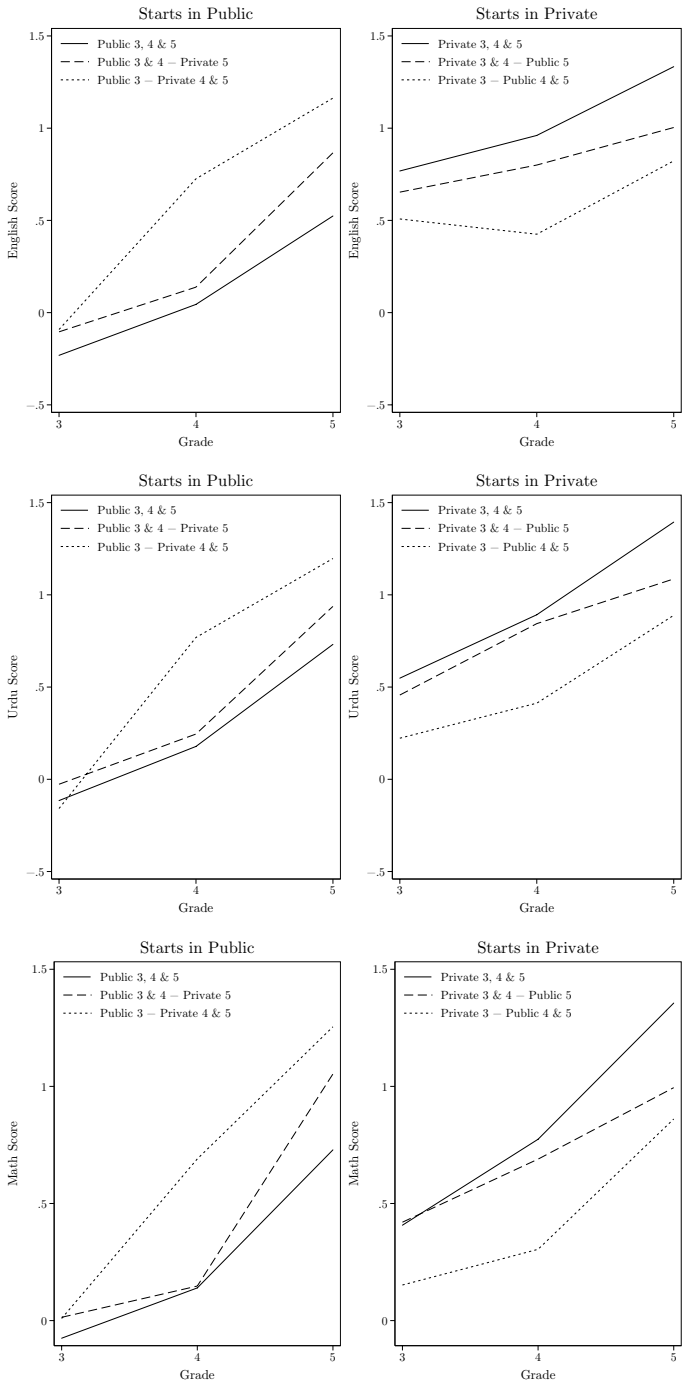
TABLE 3. THIRD GRADE ACHIEVEMENT AND CHILD, HOUSEHOLD AND SCHOOL CHARACTERISTICS

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Dependent variable (third grade):	**English**	**English**	**English**	**Urdu**	**Urdu**	**Urdu**	**Math**	**Math**	**Math**
Private School	0.985	0.907	0.916	0.670	0.595	0.575	0.512	0.446	0.451
	(0.047)***	(0.048)***	(0.048)***	(0.049)***	(0.050)***	(0.047)***	(0.051)***	(0.053)***	(0.052)***
Age		0.004	0.015		0.013	0.013		0.033	0.048
		(0.013)	(0.012)		(0.013)	(0.012)		(0.014)**	(0.013)***
Female		0.125	0.133		0.209	0.205		-0.040	-0.057
		(0.047)***	(0.041)***		(0.046)***	(0.040)***		(0.051)	(0.043)
Years of schooling		-0.029	-0.019		-0.039	-0.028		-0.038	-0.025
		(0.013)**	(0.012)		(0.014)***	(0.014)**		(0.015)**	(0.014)*
Number of elder brothers		-0.030	-0.035		-0.020	-0.025		-0.020	-0.023
		(0.011)***	(0.010)***		(0.012)*	(0.011)**		(0.012)*	(0.011)**
Number of elder sisters		0.008	0.013		0.001	-0.001		-0.002	-0.006
		(0.011)	(0.010)		(0.012)	(0.012)		(0.013)	(0.012)
Height z-score (normalized to U.S.)		0.027	0.016		0.017	0.012		0.034	0.024
		(0.007)***	(0.006)***		(0.006)***	(0.006)**		(0.008)***	(0.007)***
Weight z-score (normalized to U.S.)		-0.005	-0.001		-0.004	0.001		-0.009	-0.002
		(0.008)	(0.006)		(0.005)	(0.005)		(0.007)	(0.006)
Asset index		0.041	0.050		0.043	0.045		0.030	0.034
		(0.012)***	(0.009)***		(0.011)***	(0.010)***		(0.011)***	(0.010)***
Mother educated past elementary		0.048	0.062		0.014	0.011		0.023	-0.006
		(0.036)	(0.031)**		(0.040)	(0.035)		(0.040)	(0.037)
Father educated past elementary		0.061	0.066		0.062	0.049		0.069	0.053
		(0.033)*	(0.028)**		(0.034)*	(0.031)		(0.035)**	(0.032)*
Mother lives at home		-0.131	-0.025		-0.174	-0.108		-0.210	-0.091
		(0.095)	(0.081)		(0.102)*	(0.092)		(0.097)**	(0.090)
Father lives at home		0.006	-0.038		0.019	0.005		-0.009	-0.026
		(0.049)	(0.044)		(0.053)	(0.048)		(0.057)	(0.051)
Survey Date		0.003	0.000		0.001	0.004		0.003	0.003
		(0.002)	(0.004)		(0.002)	(0.003)		(0.002)	(0.003)
Constant	-0.243	-49.721	-3.690	-0.137	-23.750	-59.528	-0.095	-56.196	-51.248
	(0.038)***	(38.467)	(62.432)	(0.035)***	(31.915)	(45.357)	(0.038)**	(35.415)	(50.310)
Village Fixed Effects	No	No	Yes	No	No	Yes	No	No	Yes
Observations	4031	4031	4031	4031	4031	4031	4031	4031	4031
R-squared	0.23	0.25	0.37	0.11	0.13	0.25	0.06	0.08	0.21

* significant at 10%; ** significant at 5%; *** significant at 1%

Notes: Standard errors clustered at the school level. Sample includes only those children tested and surveyed in all three years.

FIGURE 3. ACHIEVEMENT OVER TIME FOR CHILDREN WHO SWITCHED SCHOOL TYPES



	Public 3, 4, & 5	Public 3, 4 – Private 5	Public 3 – Private 4 & 5	Private 3, 4 & 5	Private 3& 4 – Public 5	Private 3 – Public 4, 5
N	5688	40	48	2007	160	167

Notes: Lines connect group means for children who were enrolled in all three periods and have a particular private/public enrollment pattern. Children were tested in the second half of the school year; most of the gains from a child in a third grade government school and fourth grade private school should be attributed to the private school.

Table 4. Correcting for Measurement Error Bias

Strategy	**English**	**Urdu**	**Math**
No Correction (OLS)	0.65	0.66	0.69
	(0.015)	(0.013)	(0.013)
Alternate Subject Scores (2SLS)	0.85	0.89	0.97
	(0.018)	(0.015)	(0.019)
	[0.000]	[0.000]	[0.000]
Lagged Scores (2SLS)	0.88	0.86	0.93
	(0.019)	(0.019)	(0.020)
	[0.140]	[0.637]	[0.262]
Alternate Subjects and Lagged Scores (2SLS)	0.81	0.80	0.85
	(0.016)	(0.014)	(0.015)
	[0.000]	[0.000]	[0.000]
Analytical Correction (HEIV OLS)	0.90	0.87	0.88
	(0.020)	(0.016)	(0.017)

Notes: Cells contain coefficients from a regression of round 3 test scores on round 2 test scores—i.e. the lagged value-added model with no covariates. Parentheses contain standard errors clustered at the school level. Brackets contain the p-value for Hansen's J statistic testing the overidentifying restrictions. The 2SLS estimates use alternate subjects or lagged scores as instruments, or both. These estimators have 1, 2 and 3 overidentifying restrictions, respectively. The analytical correction uses the score standard errors from IRT to blow-up the estimate appropriately (see Appendix A). All regressions use the same set of children.

TABLE 5. VALUE-ADDED MODEL ESTIMATES OF DEPRECIATION AND PRIVATE SCHOOL COEFFICIENT (ENGLISH)

Estimator's Key Assumption	**Depreciation Coefficient**	**Private School Coefficient**	**Hansen's J χ^2 (p-value)**	**df**
Panel A: Static Estimates				
No depreciation β=*1* (OLS)	1.00	-0.08		
		(0.02)		
No effects, no measurement error (OLS)	0.52	0.31		
	(0.02)	(0.02)		
No effects (2SLS)	0.70	0.16	4.69	1
	(0.02)	(0.02)	(0.03)	
No effects (HEIV)	0.74	0.21		
	(0.02)	(0.02)		
Panel B: Difference GMM				
Strictly exogenous inputs	0.38	0.28	17.00	12
	(0.09)	(0.08)	(0.15)	
Predetermined inputs	0.41	0.54	7.00	6
	(0.09)	(0.33)	(0.32)	
Panel C: Levels and Difference SGMM				
Predetermined inputs, constantly	0.46	0.32	34.56	22
correlated effects	(0.08)	(0.07)	(0.04)	
Predetermined inputs, uncorrelated	0.50	0.38	45.27	28
effects	(0.06)	(0.04)	(0.02)	
Panel D: Levels Only GMM				
Predetermined inputs, constantly	0.47	0.25	20.32	11
correlated effects, conditional stationarity	(0.04)	(0.06)	(0.04)	
Predetermined inputs, uncorrelated	0.48	0.37	19.82	
effects, conditional stationarity	(0.05)	(0.04)	(0.03)	10

Notes: Dots represent the estimated coefficients, thicker dark lines are 90 percent confidence intervals, and thin gray lines are 95 percent confidence intervals. All intervals and standard errors are clustered by school. See text for details on instruments and assumptions.

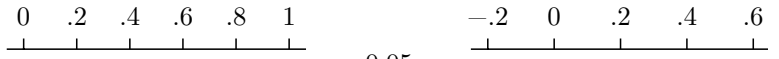
TABLE 6. VALUE-ADDED MODEL ESTIMATES OF DEPRECIATION AND PRIVATE SCHOOL COEFFICIENT (URDU)

Estimator's Key Assumption	Depreciation Coefficient	Private School Coefficient	Hansen's J χ^2 (p-value)	df
	0 .2 .4 .6 .8 1	−.2 0 .2 .4 .6		
Panel A: Static Estimates				
No depreciation β=*1* (OLS)	1.00	0.01		
		(0.02)		
No effects, no measurement error (OLS)	0.58	0.26		
	(0.01)	(0.02)		
No effects (2SLS)	0.73	0.17	3.67	1
	(0.02)	(0.02)	(0.06)	
No effects (HEIV)	0.79	0.20		
	(0.02)	(0.02)		
Panel B: Difference GMM				
Strictly exogenous inputs	0.34	0.31	49.73	12
	(0.10)	(0.08)	(0.00)	
Predetermined inputs	0.50	1.01	17.85	6
	(0.12)	(0.43)	(0.01	
Panel C: Levels and Difference SGMM				
Predetermined inputs, constantly	0.47	0.38	56.16	22
correlated effects	(0.10)	(0.07)	(0.00)	
Predetermined inputs, uncorrelated	0.49	0.35	58.17	28
effects	(0.07)	(0.05)	(0.00)	
Panel D: Levels Only GMM				
Predetermined inputs, constantly	0.56	0.31	13.65	11
correlated effects, conditional stationarity	(0.04)	(0.07)	(0.25)	
Predetermined inputs, uncorrelated	0.56	0.27	13.62	
effects, conditional stationarity	(0.04)	(0.03)	(0.19)	10

Notes: Dots represent the estimated coefficients, thicker dark lines are 90 percent confidence intervals, and thin gray lines are 95 percent confidence intervals. All intervals and standard errors are clustered by school. See text for details on instruments and assumptions.

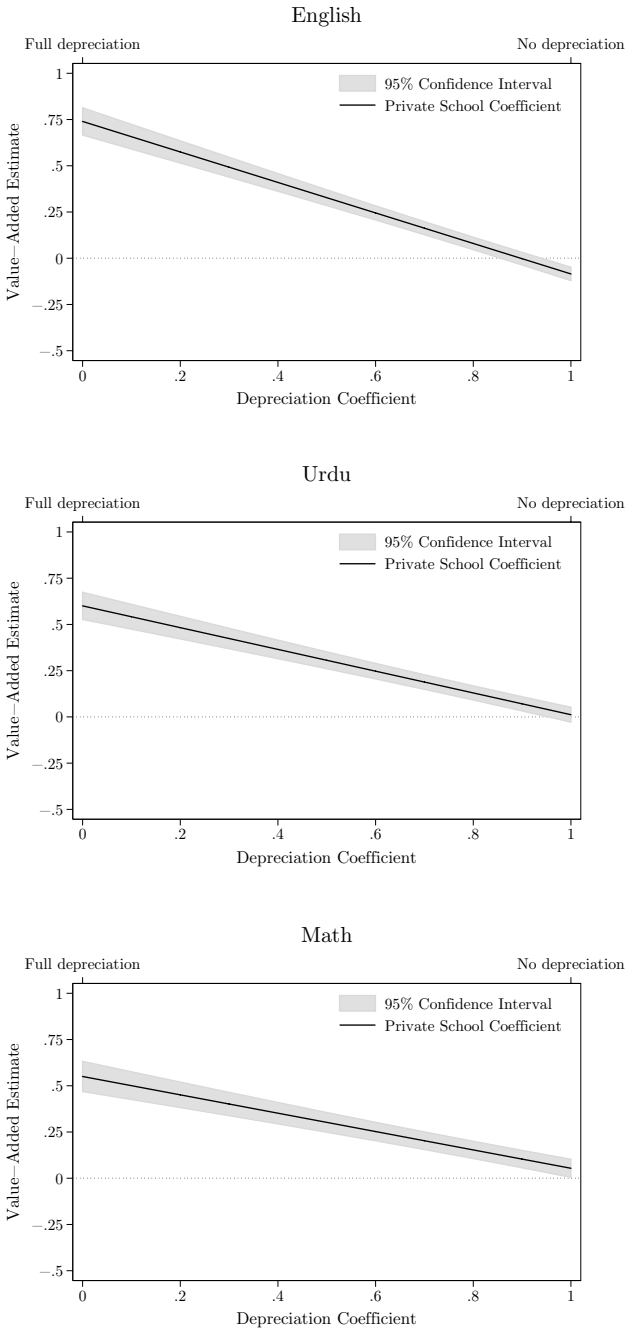
TABLE 7. VALUE-ADDED MODEL ESTIMATES OF DEPRECIATION AND PRIVATE SCHOOL COEFFICIENT (MATH)

Estimator's Key Assumption	Depreciation Coefficient	Private School Coefficient	Hansen's J χ^2 (p-value)	df
	0 .2 .4 .6 .8 1	−.2 0 .2 .4 .6		
Panel A: Static Estimates				
No depreciation β=*1* (OLS)	1.00	0.05		
		(0.02)		
No effects, no measurement error (OLS)	0.57	0.27		
	(0.02)	(0.03)		
No effects (2SLS)	0.76	0.17	0.02	1
	(0.02)	(0.03)	(0.89)	
No effects (HEIV)	0.75	0.23		
	(0.02)	(0.03)		
Panel B: Difference GMM				
Strictly exogenous inputs	0.09	0.28	29.25	12
	(0.07)	(0.10)	(0.00)	
Predetermined inputs	0.12	0.37	14.68	6
	(0.07)	(0.39)	(0.02)	
Panel C: Levels and Difference SGMM				
Predetermined inputs, constantly	0.19	0.35	48.94	22
correlated effects	(0.07)	(0.09)	(0.00)	
Predetermined inputs, uncorrelated	0.25	0.44	59.04	28
effects	(0.07)	(0.04)	(0.01	
Panel D: Levels Only GMM				
Predetermined inputs, constantly	0.58	0.33	31.23	11
correlated effects, conditional stationarity	(0.04)	(0.07)	(0.00)	
Predetermined inputs, uncorrelated	0.59	0.27	30.35	
effects, conditional stationarity	(0.04)	(0.04)	(0.00)	10



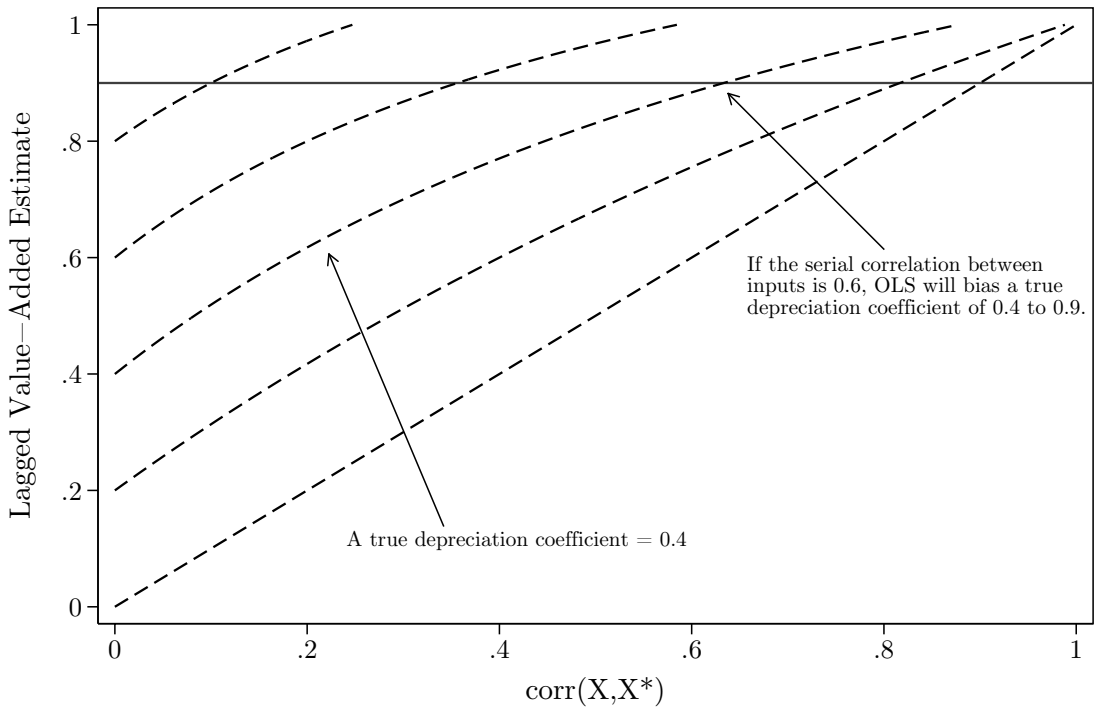
Notes: Dots represent the estimated coefficients, thicker dark lines are 90 percent confidence intervals, and thin gray lines are 95 percent confidence intervals. All intervals and standard errors are clustered by school. See text for details on instruments and assumptions.

FIGURE 4. PRIVATE SCHOOL VALUE-ADDED ASSUMING VARIOUS DEPRECIATION RATES



Notes: These graphs show the estimated value-added effect of private schools depending on the assumed depreciation coefficient of lagged achievement. The restricted value-added model, for example, assumes the depreciation coefficient equals one—no depreciation. The estimated value-added pooled for third to fourth and fourth to fifth grades is estimated by OLS controlling for age, gender, years of schooling, weight z-score, height z-score, number of elder brothers, number of elder sisters, whether father lives at home, whether mother lives at home, whether father educated past elementary, whether mother educated past elementary, an asset index, survey date, and round and village fixed effects. The confidence intervals are based on standard errors clustered at the school level.

FIGURE 5. TRUE AND ESTIMATED DEPRECIATION IN A LAGGED VALUE-ADDED MODEL WITH SERIALLY CORRELATED OMITTED INPUTS



Notes: In the lagged value-added model the depreciation coefficient is biased upward by the correlation between omitted contemporaneous inputs and past inputs that are captured in the lagged test score. Assuming constant correlation between any two years of inputs X and X* the bias can be calculated analytically (see text). The graph above gives the implied bias for fourth grade. The dashed lines represent the true depreciation coefficient, indicated by the associated β when corr(X,X*)=0. The y-axis gives the biased estimate that results from estimating a lagged value-added model. This estimate depends on the true depreciation rate (dashed lines) and the assumed correlation of inputs over time (x-axis). For example, a (biased) estimated depreciation coefficient of 0.9 may result from a true depreciation coefficient of 0.4 and correlation between inputs around 0.6. These calculations assume achievement is measured perfectly and all inputs are omitted (i.e. unobserved) in the regression.

Table 7. Depreciation and Equal Selection on Observed and Unobserved variables

	English	**Urdu**	**Math**
Lagged Score (no controls)	0.91	0.92	0.89
R-Squared (no controls)	0.52	0.52	0.46
Lagged Score (with controls)	0.74	0.79	0.75
R-Squared (with controls)	0.56	0.56	0.52
Implied remaining bias	0.72	0.77	0.75
Bias corrected estimate	0.03	0.02	0.00

Notes: The lagged value-added model with controls includes a private school dummy, age, gender, years of schooling, weight z-score, height z-score, number of elder brothers, number of elder sisters, whether father lives at home, whether mother lives at home, whether father educated past elementary, whether mother educated past elementary, an asset index, survey date, and round and village fixed-effects.

The implied remaining bias calculation is derived from the assumption of equal selection on observables and unobservable, introduced by Altonji et al. (2005). The bias corrected estimate is the lagged score estimate (with controls) minus the estimated bias. The implied bias and bias corrected estimate should be viewed as bounds. We modify the equal selection on observed and unobserved variables assumption to incorporate the fact that the fraction of explainable variance is less than one (roughly 0.9), due to measurement error.

TABLE 8. HOUSEHOLD RESPONSES TO PERFORMANCE SHOCKS

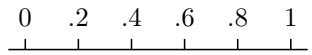
	Test Score Residual (Grade 4)			
Household Changes (Grade 4 to 5)	**English**	**Urdu**	**Math**	**N**
Perception of child performance	0.25***	0.17**	0.20***	652
	(0.09)	(0.07)	(0.08)	
Log expenditure on school	-0.05	-0.07*	-0.03	643
	(0.05)	(0.04)	(0.04)	
Hours helping child	-0.66	-0.34	-0.44	645
	(0.46)	(0.36)	(0.38)	
Log minutes spent on homework	-0.14	0.10	0.04	617
	(0.25)	(0.20)	(0.21)	
Log minutes for tuition	0.21	0.10	0.02	620
	(0.20)	(0.16)	(0.17)	
Log minutes spent playing	0.54*	0.27	0.29	619
	(0.30)	(0.23)	(0.25)	

* significant at 10%; ** significant at 5%; *** significant at 1%

Notes: The grade 4 test score residual is computed from a lagged value-added model OLS regression that controls for third grade scores and a comprehensive set of household controls (age, gender, health status, household size, elder brothers, elder sisters, father education, mother education, adult education index, minutes spent helping child, asset index, log monthly expenditure, and wealth relative to village) . The coefficients and standard errors reported are for separate 2SLS regressions of the grade 4 to 5 household behavior change on the subject residual, instrumented using the alternate subject residuals. Roughly half of the households received the score results as part of a randomized evaluation of school and child report cards. Logged variables are computed as ln(1+x).

TABLE 9. DEPRECIATION COEFFICIENT USING WITHIN AND BETWEEN SCHOOL VARIATION ONLY

Subject	Variation	Depreciation Coefficient
English	Within	0.57
		(0.02)
	Between	0.45
		(0.09)
Urdu	Within	0.64
		(0.02)
	Between	0.31
		(0.10)
Math	Within	0.64
		(0.03)
	Between	0.13
		(0.14)



Notes: Depreciation coefficient are calculated using a 2SLS regression of test scores on lagged test scores, instrumented using lagged differences in alternate subjects (i.e. basic moments from conditional stationarity). Within regressions use school demeaned child scores whereas between regressions use mean school scores. The sample is from Table 2, Panel A with no covariates. Within N = 8620, between N = 761.

TABLE 10. EXPERIMENTAL ESTIMATES OF PROGRAM FADE OUT

Program	Subject	Immediate Treatment Effect	Extended Treatment Effect	Implied Depreciation Coefficient (1-δ)	Source
Balsakhi Program	Math	0.348	0.030	0.086	Banerjee et al
	Verbal	0.227	0.014	0.062	(2007)
CAL Program	Math	0.366	0.097	0.265	Banerjee et al
	Verbal	0.014	-0.078	~0.0	(2007)
Learning Incentives	Multi-subject	0.23	0.16	0.70	Kremer et al (2003)
Teacher Incentives	Multi-subject	0.139	-0.008	~0.0	Glewwe et al (2003)
STAR Class Size Experiment	Stanford-9 and CTBS	~5 percentile points	~2 percentile points	~ .25 to .5	Krueger and Whitmore (2001)
Summer School and Grade Retention	Math	0.136	0.095	0.70	Jacob and Lefgren
	Reading	0.104	0.062	0.60	(2004)

Notes: Extended treatment effect is achievement approximately one year after the treatment ended. Unless otherwise noted, effects are expressed in standard deviations. Results for Kremer et al. (2003) are averaged across boys and girls. Estimated effects for Jacob and Lefgren (2004) are taken for the third grade sample.

Table 11. Depreciation Coefficient Heterogeneity Across School, Child and Household Characteristics

	Depreciation Coefficient		
Within category:	**English**	**Urdu**	**Math**
	0 .2 .4 .6 .8 1	0 .2 .4 .6 .8 1	0 .2 .4 .6 .8 1
Private School	0.36	0.48	0.38
	(0.07)	(0.07)	(0.08)
Public School	0.53	0.58	0.62
	(0.05)	(0.05)	(0.05)
Female	0.35	0.61	0.56
	(0.08)	(0.05)	(0.06)
Male	0.47	0.53	0.54
	(0.06)	(0.05)	(0.06)
Richer Family	0.37	0.49	0.47
	(0.08)	(0.07)	(0.07)
Poorer Family	0.51	0.56	0.58
	(0.08)	(0.06)	(0.07)
Mother educated past primary	0.33	0.54	0.40
	(0.12)	(0.08)	(0.09)
Mother not education past primary	0.49	0.58	0.63
	(0.06)	(0.05)	(0.05)
Father educated past primary	0.48	0.61	0.55
	(0.07)	(0.05)	(0.05)
Father not educated beyond primary	0.44	0.55	0.59
	(0.07)	(0.05)	(0.06)

Notes: This table reports estimates for specific sub-populations; each coefficient is from a separate regressions. The coefficients are estimated using 2SLS (levels only) under the assumption of predetermined inputs, constantly correlated effects and conditional stationarity. Standard errors are clustered at the school level.