Here Today, Gone Tomorrow? Examining the Extent and Implications of Low Persistence in Child Learning

Tahir Andrabi   Jishnu Das   Asim Ijaz Khwaja   Tristan Zajonc[*]

First Draft: April 18, 2007[†]
This Draft: January 2, 2009

Abstract

Learning persistence plays a central role in models of skill formation, estimates of education production functions, and evaluations of educational programs. In non-experimental settings, estimated impacts of educational inputs can be highly sensitive to correctly specifying persistence when inputs are correlated with baseline achievement. While less of a concern in experimental settings, persistence still links short-run treatment effects to long-run impacts. We study learning persistence using dynamic panel methods that account for two key empirical challenges: unobserved student-level heterogeneity in learning and measurement error in test scores. Our estimates, based on detailed primary panel data from Pakistan, suggest that only a fifth to a half of achievement persists between grades. Using private schools as an example, we show that incorrectly assuming high persistence significantly understates and occasionally yields the wrong sign for private schools' impact on achievement. Towards an economic interpretation of low persistence, we use question-level exam responses as well as household expenditure and time-use data to explore whether psychometric testing issues, behavioral responses, or forgetting contribute to low persistence—causes that have different welfare implications.

JEL Classifications: I21, J24, C23, O12, H4,
Keywords: education, fade out, learning persistence, value-added models, dynamic panel data, private schools.

1 Introduction

Models of learning often assume that children's achievement persists between grades—what a child learns today largely stays with her tomorrow. Yet recent research highlights that treatment effects measured by test scores fade rapidly, both in randomized interventions and observational studies. Jacob, Lefgren and Sims (2008), Kane and Staiger (2008), and Rothstein (2008) find that teacher effects dissipate by between 50 and 80 percent over one year. The same pattern holds in several studies of supplemental education programs in developed and developing countries. Currie and Thomas (1995) document the rapid fade out of Head Start's impact in the United States, and Glewwe, Ilias and Kremer (2003) and Banerjee et al. (2007) report on education experiments in Kenya and India where over 70 percent of the one-year treatment effect is lost after an additional year. Low persistence may in fact be the norm rather than the exception, and a central feature of learning.

Low persistence has broad implications. Commonly used empirical methods often make strong assumptions about persistence that can affect both short-run impact estimates and long-run extrapolations. Moreover, to the degree that low persistence is not an artifact of psychometric issues, such as measurement error, changing test content, or cheating, its extent and interpretation also matters. If low persistence arises from economic responses, such as parents and teachers devoting fewer resources to better performers, such subsitution may lead to total costs falling enough that welfare actually rises, even when achievement gains fade out. However, if low persistence arises from biological factors, such as the inherent fragility of human memory, welfare likely suffers due to an unavoidable inefficiency in learning.

This paper studies the extent, implications, and interpretation of learning persistence using primary panel data from Pakistan that follows over eight thousand children in public and private primary schools and includes information on learning outcomes matched with household educational inputs.[1] There are four broad findings. First, our estimates show that learning persistence is low: only a fifth to a half of achievement persists between grades. Second, in estimating the degree of persistence, we find significant biases arise both due to unobserved student-level heterogeneity in learning and measurement error in test scores. We attempt to address both: the former, by applying a variety of dynamic panel data techniques typically used in the economic growth literature; the latter, by instrumenting with alternate subject test scores and by analytically correcting for test measurement error using reliability estimates obtained from Item Response Theory.[2]

---

[1] There are several different uses of the term persistence in the education literature. We refer to persistence as the fraction of knowledge that persist from one period to the next. The education literature, however, also uses the term "persistence" to indicate the probability of continuation from grade to grade (as opposed to dropping out), or to indicate a child's motivation or propensity to complete tasks in the face of adversity.

[2] 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

Third, we use the example of private school attendance to illustrate that estimates of educational inputs' learning impacts can be highly sensitive to the persistence parameter. This is particularly true for inputs that are continually applied and lead to a large baseline gap in achievement, as is the case with our specific application. We find that incorrectly assuming high persistence significantly understates and occasionally yields the wrong sign for private schools' impact on achievement.

Finally, we use question-level test data as well as detailed household expenditure and time-use data to shed some light on the possible factors behind low persistence. We find little support for measurement error, mechanical psychometric testing, and behavioral response explanations. This suggests that forgetting may be an important factor.

Figures 1 and 2 provide a simple illustration of low persistence and its potential implications. Figure 1 plots learning levels in the tested subjects (English, mathematics, and the vernacular, Urdu) for roughly eight thousand children in public and private schools in Pakistan over three years. While, levels are always higher for children in private schools, there is little difference in learning gains (the gradient) between public and private schools. Therefore, a specification that uses learning gains (i.e., assumes perfect persistence) would conclude that private schools add no greater value to learning than their public counterparts. Figure 2 reverses this conclusion. It instead directly plots learning gains in public and private schools against a child's *initial* learning level. At every point of the initial score distribution, children in private schools gain more than children in public schools. This reversal from Figure 1 occurs because, as is apparent in Figure 2, (i) learning gains decrease with higher initial scores—consistent with only a fraction of achievement persisting between grades[3]—and (ii) children in private school start off at higher levels—as indicated by the kernel densities in Figure 2. In this situation, comparing learning gains without conditioning on initial learning differences (Figure 1) can be potentially misleading since it does not take into account initial differences in learning between the two groups.

Figures 1 and 2 are graphical representations of two popular value-added specifications of education production functions used to estimate program effects in observational studies. Figure 1 is the analog of the restricted value-added or gain-score model. This model assumes that lagged achievement contributes cumulatively *without loss* to future achievement. The effect of a school or a teacher who spurs a child a few months ahead, say, will persist in third, fourth, and even twelfth grade. By comparison, Figure 2 represents the more flexible lagged

---

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.

[3]The persistence parameter is modeled as the *fraction* (not absolute level) of initial learning that persists. Low persistence therefore implies that, conditional on the same input, initially higher performers will gain less. In contrast, under perfect persistence implies that learning gains are not a function of initial learning levels.

value-added model where the contribution of previous achievement may decay. Clearly, the choice of specification matters. Because learning gains are characterized by low persistence, bias or misspecification in the estimation of persistence can bias the estimate of the relevant input, in this case, the contribution of private schooling to learning.

While Figure 2 offers a more accurate picture, estimating the associated lagged value-added model poses two main difficulties. First, in specifications that include lagged achievement as an explanatory variable, unobserved student heterogeneity that speeds learning (i.e., affects learning rates not just achievement levels) biases the estimated coefficient on lagged achievement upward. Second, test scores are characterized by measurement error. Measurement error counteracts the upward heterogeneity bias by attenuating the coefficient on lagged achievement.

We address these issues using dynamic panel methods that can account for unobserved heterogeneity in learning together with an additional correction for measurement error. Unobserved heterogeneity in learning is analogous to the concerns that arise in controlling for initial income levels in the economic growth literature and has spurred a rich literature in dynamic panel data methods (e.g. Arellano and Honore, 2001; Arellano, 2003). The most common approach uses the twice-lagged scores as an instrumental variable in a differenced model (Arellano and Bond, 1991). However, differencing may be problematic if there is little variation over time in the relevant input, as is common in many educational applications, or if the model has a unit root, as perfect persistence implies. We therefore also explore versions of "system" GMM estimators that have better properties when persistence approaches one and can account for unobserved heterogeneity while allowing for the estimation of time-invariant inputs that might otherwise be unidentifiable using more common difference GMM methods (Arellano and Bover, 1995). To address measurement error, we instrument using alternate subjects or use an analytical correction derived from Item Response Theory. Given that we do have some time-series input variation in the use of public and private schools—5 percent of children switch schools per year—and persistence is low, we find that both difference and system GMM estimators that correct for measurement error and unobserved heterogeneity yield relatively similar results.

Together, these techniques allow us to provide improved estimates of learning persistence in a model that allows for imperfect persistence, unobserved heterogeneity in learning, and measurement error in test scores. Moreover, we demonstrate the importance of estimating the persistence parameter correctly on educational inputs of interest, in our case, the impact of attending private schools. To elaborate further on our findings, our preferred estimates suggest that the coefficient on lagged achievement in the subjects of English, Urdu, and mathematics is between 0.2 and 0.5, with some estimators yielding significantly lower estimates. Furthermore, there is evidence of both omitted learning heterogeneity—some students learn faster—and measurement error. Correcting only for measurement error yields estimates of the persistence parameter between 0.7 and 0.79. This raises a note of caution: correcting for mea-

surement error alone may be worse than doing nothing.[4] As is then expected, the estimated contribution of an educational input is highly sensitive to the persistence parameter when the baseline achievement gap is large. Assuming perfect persistence suggests that private schools contribute no more than public schools; in contrast, the dynamic panel estimates suggest large and significant contributions ranging from 0.19 to 0.32 standard-deviations a year.[5]

Finally, we also present preliminary evidence on the the economic interpretation of persistence. Using detailed data on matched household-school pairs available in our survey, we do not find strong evidence that parents substitute away from children who receive a positive achievement shock. While the coefficients' signs generally support a substitution rather than reinforcement effect, the size of these effects are small and marginally significant, if at all. There is also little evidence that low persistence arises because teachers target struggling children; persistence within classrooms, where targeting might occur, is as high as that across classrooms. Neither are mechanical explanations, such as "teaching to the test" or cheating, likely responsible for low estimated persistence. Our test was designed to cover standard curricular knowledge, was relatively low-stakes, and was administered directly by our survey team to prevent cheating. Our results are also not driven by changes in the exam content; limiting scores to common questions administered across multiple years yields similar results. A final explanation, consistent with psychological evidence from other studies, suggests that forgetting may be a key explanation of our results. This area remains ripe for further research.

We conclude by discussing the implications of low persistence for non-experimental and experimental studies. In non-experimental settings, given three periods of data, dynamic panel methods can recover both the effect of educational inputs and the persistence parameter. With only two years of data, assuming perfect persistence (the gain-score model) can lead to highly misleading estimates, particularly when baseline gaps are large. In this case, the lagged value-added model performs significantly better for estimating inputs' impacts by comparing children with similar baseline characteristics. While comparing children with similar baseline characteristics is an intuitive approach for estimating a program's impact, the same logic does not apply for estimating the persistence parameter. In the lagged value-added model, the persistence

---

[4]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.

[5]Harris and Sass (2006) find the that the persistence parameter makes little difference to estimates of teacher effects, while we find it starkly affects the estimates of school type. This can be explained by the relative gaps in baseline achievement. It is likely that a child does not continue with the same teacher, or an equally good teacher, over time. Hence, even if we don't observe children's educational history, two children who currently have different teachers may have been exposed to a similar quality teachers in the past. As such, children with different teachers often do not differ substantially in their baseline learning levels. In contrast, given that there is little switching across school types, children currently in different schools often differ substantially in baseline learning levels, as evidenced in Figures 1 and 2.

parameter remains biased by unobserved heterogeneity and measurement error attenuation. Without a consistent estimate of persistence, extrapolating long-run effects from short-run analysis in both non-experimental and experimental setups is difficult.[6] Moreover, the interpretation of persistence also matters: household substitution, targeting or tracking, forgetting, and psychometric concerns have fundamentally different implications for calculating the full costs and benefits of particular programs.

Three recent papers address concerns related to the ones we study, albeit in the context of estimating teacher value-added effects. Rothstein (2008) tests dynamic student tracking in value-added models. He finds rapid fade-out of teacher effects and that short-run effects are only weakly correlated with longer-run effects. Jacob, Lefgren and Sims (2008) study the persistence of teacher effects and propose estimating an input's persistence by instrumenting using exogenous lagged treatment assignments. Kane and Staiger (2008) validate value-added models by comparing observational value-added teacher effects with their experimental counterparts, using a sample of experimental teacher assignments from the Los Angeles Unified School District. All three studies corroborate our findings of low persistence in test scores; in particular, Jacob, Lefgren and Sims (2008) reports one year persistence rates around 0.2, consistent with our results.[7]

Our context is somewhat different, not only because we consider an emerging economy and a different input, school type, but because the nature of the input is such that a child is more likely to experience the same input over time. This context more naturally lends itself to examining how sensitive the impact of the educational input is to correctly estimating persistence. Moreover, since we utilize detailed primary data from an extensive educational project that we designed, we can explore mechanical and psychometric explanations for low persistence using item data from a testing environment we directly controlled. We can also explore behavioral explanations using rich matched household input and time-use data. Lastly, low persistence is quite striking in our context given that the low learning levels we find likely imply that tests scores are closely aligned with basic educational goals.[8]

---

[6]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). Although effects on learning as measured by test-scores may fade, non-cognitive skills that are rewarded in the labor market could persist. For instance, Currie and Thomas (1995), Deming (2008) and Schweinhart et al. (2005) provide evidence of long run effects of Head Start and the Perry Preschool Project, even though cognitive gains largely fade after children enroll in regular classes.

[7]The idea that knowledge fades over time has studied carefully in psychology for over one hundred years (e.g. Ebbinghaus, 1885; Rubin and Wenzel, 1996). In value-added models, it dates back to at least Boardman and Murnane (1979). More recently, both Kane and Staiger (2002) and Chay, McEwan and Urquiola (2005) highlight the implications of mean reverting test scores for specific applications.

[8]On average, the children tested at the end of Grade 3 could complete two-digit addition, but not subtraction or multiplication (in mathematics); recognize simple words (but not sentences) in the vernacular (Urdu); and recognize alphabets and match simple three-letter words to pictures in English.

A final contribution of our work is that it applies a wider set of econometric tools from the dynamic panel data literature than have been typically used in the education literature. In the use of dynamic panel methods, our estimators bear greatest resemblance to those discussed by Schwartz and Zabel (2005) and Sass (2006). Both use simple dynamic panel estimators, in the first case using school-level data and in the second using the Arellano and Bond (1991) differences GMM approach. Santibanez (2006) also uses the Arellano and Bond (1991) estimator to analyze the impact of teacher quality. Our efforts extend to include system GMM approaches and to address measurement error in test scores and alternative assumptions regarding omitted heterogeneity.

The rest of the paper is organized as follows: Section 2 presents the basic education production function analogy and discusses the specification and estimation of the value-added approximations to it. Section 3 summarizes our data. Section 4 reports our main results and several robustness checks. Section 5 provides a preliminary exploration of the economic interpretation of persistence. Section 6 concludes by discussing implications for experimental and non-experimental program evaluation.

2   Empirical Learning Framework

The "education production function" approach to learning relates current achievement to all previous inputs. Boardman and Murnane (1979) and Todd and Wolpin (2003) provide two excellent accounts of this approach and the assumptions it requires; the following is a brief summary.[9] Using notation consistent with the dynamic panel literature, we aggregate all inputs into a single vector $\mathbf{x}_{it}$ and 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_{it}^* is true achievement, measured without error, and the summed μ_{is} are cumulative productivity shocks.[10] Estimating (1) is generally impossible because researchers do not observe

---

[9]Researchers generally assume that the model is additively separable across time and that input interactions can be captured by separable linear interactions. Cunha, Heckman and Schennach (2006) and Cunha and Heckman (2007) are two exceptions to this pattern, where dynamic complementarity between early and late investments and between cognitive and non-cognitive skills are permitted.

[10]This starting point is more restrictive than the more general starting framework presented by Todd and Wolpin (2003). In particular, it assumes an input applied in first grade has the same effect on first grade scores as an input applied in second grade has on second grade scores.

the full set of inputs, past and present. The value-added strategy makes estimation feasible by rewriting (1) to avoid the need for past inputs. Adding and subtracting βy_{it}^*, normalizing θ_1 to unity, and assuming that coefficients decline geometrically ($\alpha_j = \beta \alpha_{j-1}$ and $\theta_j = \beta \theta_{j-1}$ for all j) yields the *lagged value-added model*

$$y_{it}^* = \boldsymbol{\alpha}_1' \mathbf{x}_{it} + \beta y_{i,t-1}^* + \mu_{it}. \tag{2}$$

The basic idea behind this specification is that lagged achievement will capture the contribution of all previous inputs and any past unobservable endowments or shocks. We refer to α as the *input coefficient* and β as the *persistence coefficient*. Finally, imposing the restriction that $\beta = 1$ yields the gain-score or *restricted value-added model* that is often used in the education literature:

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

This model asserts that past achievement contains no information about future gains, or equivalently, that an input's effect on any subsequent level of achievement does not depend on how long ago it was applied. As we will see from our results, the assumption that $\beta = 1$ is clearly violated in the data, and increasingly it appears, in the literature, as well. As a result, we will focus primarily on estimating (2).

There are two potential problems with estimating (2). First, the error term μ_{it} could include individual (child-level) heterogeneity in *learning* (e.g., $\mu_{it} \equiv \eta_i + \upsilon_{it}$). Lagged achievement only captures individual heterogeneity if it enters through a one-time process or endowment, but talented children may also *learn* faster. Since this unobserved heterogeneity enters in each period, $\text{Cov}(y_{i,t-1}^*, \mu_{it}) > 0$ and β will be biased upwards.

The second likely problem is that test scores are inherently a noisy measure of latent achievement. Letting $y_{it} = y_{it}^* + \varepsilon_{it}$ denote observed achievement, we can rewrite the latent lagged value-added model (2) 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 persistence coefficient, the expected bias due to both of these sources is

$$\text{plim} \, \beta_{OLS} = \beta + \left(\frac{\text{Cov}(\eta_i, y_{i,t-1}^*)}{\sigma_{y^*}^2 + \sigma_\varepsilon^2} \right) - \left(\frac{\sigma_\varepsilon^2}{\sigma_{y^*}^2 + \sigma_\varepsilon^2} \right) \beta. \tag{3}$$

The coefficient is biased upward by learning heterogeneity and downward by measurement error. These effects only cancel exactly when $\text{Cov}(\eta_i, y_{i,t-1}^*) = \sigma_\varepsilon^2 \beta$ (Arellano, 2003).

Furthermore, bias in the persistence coefficient leads to bias in the input coefficients, α. To

see this, consider imposing a biased $\hat{\beta}$ and estimating the resulting model

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

The error term now includes $(\beta - \hat{\beta})y_{i,t-1}$. Since inputs and lagged achievement are generally 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 on the potential correlation between inputs and learning heterogeneity that remains in μ_{it}.

This is more clearly illustrated in the case of the restricted value-added model (assuming that $\beta = 1$) where:

$$\text{plim } \hat{\alpha}_{OLS} = \alpha - (1 - \beta)\frac{\text{Cov}(\mathbf{x}_{it}, y_{i,t-1})}{\text{Var}(\mathbf{x}_{it})} + \frac{\text{Cov}(\mathbf{x}_{it}, \eta_i)}{\text{Var}(\mathbf{x}_{it})}. \tag{4}$$

Therefore, if indeed there is perfect persistence as assumed and if inputs are uncorrelated with η_i, OLS yields consistent estimates of the parameters α. However, if $\beta < 1$, OLS estimation of α now results in two competing biases. By assuming an incorrect persistence coefficient we leave a portion of past achievement in the error term. This misspecification biases the input coefficient downward by the first term in (4). The second term captures possible correlation between current inputs and omitted learning heterogeneity. If there is none, then the second term is zero, and the bias will be unambiguously negative.

2.1 Addressing Child-Level Heterogeneity: Dynamic Panel Approaches to the Education Production Function

Dynamic panel approaches can address omitted child-level heterogeneity in value-added approximations of the education production function. We interpret the value-added model (2) as an autoregressive dynamic panel model with unobserved student-level effects:

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

$$\mu_{it} \equiv \eta_i + \upsilon_{it}. \tag{6}$$

Identification of β and α is achieved by imposing appropriate moment conditions. Following Arellano and Bond (1991) and Arellano and Bover (1995), we focus on linear moment conditions and split our analysis into three groups: "differences" GMM, "differences and levels" GMM, and "levels only" GMM, which respectively refer to whether the estimates are based on the undifferenced "levels" equation (5), a differenced equation (see equation (7) below), or both.

9

The section below provides a brief overview of the estimators we explore. Table 1 summarizes the estimators, including the standard static value-added estimators (M1-M4) and dynamic panel estimators (M5-M10). For more complete descriptions, Arellano and Honore (2001) and Arellano (2003) provide excellent reviews of these and other panel models. While dynamic panel methods have now been applied to many empirical domains, only the most basic have received any application in education. We review a wider range of possible estimators and compare their appropriateness in the education context.

2.1.1 Differences GMM: Switching estimators

As noted previously, the value-added model differences out omitted endowments that might be correlated with the inputs. It does not, however, difference out heterogeneity that speeds learning. To accomplish this, the basic intuition behind the Arellano and Bond (1991) difference GMM estimator is to difference again. Differencing the dynamic panel specification of the lagged value-added model (5) yields

$$y_{it}^* - y_{i,t-1}^* = \boldsymbol{\alpha}'(\mathbf{x}_{it} - \mathbf{x}_{i,t-1}) + \beta(y_{i,t-1}^* - y_{i,t-2}^*) + [v_{it} - v_{i,t-1}]. \tag{7}$$

Here, the differenced model eliminates the unobserved fixed effect η_i. However, (7) cannot be estimated by OLS because $y_{i,t-1}^*$ is correlated by construction with $v_{i,t-1}$ in the error term. 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 exogeneity conditions. These lags are uncorrelated with the error term but are correlated with the change in lagged achievement, provided $\beta < 1$. The input coefficient, in our case the added contribution of private schools, is primarily identified from the set of children who switch schools in the observation period.

The implementation of the difference GMM approach depends on the precise assumptions about inputs. We consider two candidate assumptions: strictly exogenous inputs (M5) and predetermined inputs (M6). 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 also consider the weaker case where inputs are predetermined but not strictly exogenous. Specifically, the predetermined inputs case assumes that inputs are uncorrelated with present and future disturbances but are potentially correlated with past disturbances. This case also assumes lagged achievement is uncorrelated with present and future disturbances. Compared to strict exogeneity, this approach uses only lagged inputs as instruments. Switching schools is instrumented by the original school type, allowing switches to depend on previous shocks. This estimator remains consistent if a child switches school at

the same time as an achievement shock but still rules out parents anticipating and adjusting to future expected shocks.

2.1.2 Levels and differences GMM: Uncorrelated or constantly correlated effects

One difficulty with the differences GMM approach (M5 and M6) is that time-invariant inputs drop out of the estimated equation and their effects are therefore not identified. In our case, this means that the identification of the private school effect is based on the five percent of children who switch between public and private schools. We address the limited time-series variation using the levels and differences GMM framework proposed by Arellano and Bover (1995) and extended by Blundell and Bond (1998). Levels and differences GMM estimates a system of equations, one for the undifferenced levels equation (5) and another for the differenced equation (7). Further assumptions regarding the correlation between inputs and heterogeneity (though not necessarily between heterogeneity and lagged achievement) yield additional instruments.

We first consider predetermined inputs that have a constant correlation with the individual effects (M7). While inputs may be correlated with the omitted effects, constant correlation implies switching is not. The constant correlation assumption implies that $\triangle \mathbf{x}_{it}$ are available as instruments in the levels equation (Arellano and Bover, 1995). In the context of estimating school type, this estimator can be viewed as a levels and differences switching estimator since it relies on children switching school types in both the levels and differences equations. In practice, we often must assume that any time-invariant inputs are uncorrelated with the fixed effect or the levels equation, which includes the time-invariant inputs, is not fully identified.

A second possibility is that inputs are predetermined but are also uncorrelated with the omitted effects (M8). This allows using inputs $\mathbf{x}_i^t$ as instruments in the levels model (5). The required assumption is fairly strong; it is natural to believe that inputs are correlated with the omitted effect. Certainly, the decision to attend private school may be correlated with the child's ability to learn. At the same time, the assumption is weaker than OLS estimation of lagged value-added model since the model (M8) allows for the omitted fixed effect to be correlated with lagged achievement.

2.1.3 Levels GMM: Conditional mean stationarity

In some instances, it may be reasonable to assume that, while learning heterogeneity exists, it does not affect achievement gains. A talented child may be so far ahead that imperfect persistence cancels the benefit of faster learning. That is, individual heterogeneity may be uncorrelated with gains, $y_{it}^* - y_{it-1}^*$, but not necessarily with *learning,* $y_{it}^* - \beta y_{it-1}^*$. This situation arises when the initial conditions have reached a convergent level with respect to the fixed effect

such that

$$y_{i1}^* = \frac{\eta_i}{1 - \beta} + d_i, \qquad (8)$$

where $t = 1$ is the first observed period and 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 on the full set of controls and common time dummies, the individual effect does not influence achievement gains.

While this assumption seems too strong in the context of education, we discuss it because the dynamic panel literature has documented large downward biases of other estimators when the instruments are weak (e.g. Blundell and Bond, 1998). This occurs when persistence is perfect ($\beta = 1$) since the lagged value-added model then exhibits a unit root and lagged tests scores become weak instruments in the differenced model. The conditional mean stationarity assumption provides an additional $T - 2$ non-redundant moment conditions that can augment the 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. In this simpler approach, we instrument the undifferenced value-added model (5) using lagged changes in achievement, $\triangle y_i^{*t-1}$, and either changes in inputs, $\triangle \mathbf{x}_i^t$, or inputs directly, $\mathbf{x}_i^t$, depending on whether inputs are constantly correlated (M9) or are uncorrelated with the individual effect (M10).

2.2 Addressing Measurement Error in Test Scores

Measurement error attenuates the coefficient on lagged achievement and can bias the input coefficient in the process. Dynamic panel estimators do not address measurement error on their own. For instance, if we replace true achievement with observed achievement in the standard Arellano and Bond (1991) setup, (7) becomes

$$\triangle y_{it} = \boldsymbol{\alpha}' \triangle \mathbf{x}_{it} + \beta \triangle y_{i,t-1} + [\triangle v_{it} + \triangle \varepsilon_{i,t} - \beta \triangle \varepsilon_{i,t-1}]. \qquad (9)$$

The standard potential instrument, $y_{i,t-2}$, is uncorrelated with $\triangle v_{it}$ but is correlated with $\triangle \varepsilon_{i,t-1} = \varepsilon_{i,t-1} - \varepsilon_{i,t-2}$ by construction.

The easiest solution is to use either three-period lagged test scores or alternate subjects as

instruments. In the dynamic panel models discussed above, correcting for measurement error using additional lags requires four years of data for each child—a difficult requirement in most longitudinal datasets, including ours. We therefore use alternate subjects, although doing so does not address the possibility of correlated measurement error across subjects.

An alternative to instrumental variables strategies is to correct for measurement error analytically using the standard error of each test score, available from Item Response Theory.[11] Because the standard error is heteroscedastic—tests discriminate poorly between children at the tails of the ability distribution—one can gain efficiency by using the heteroscedastic errors-in-variables (HEIV) procedure outlined in Sullivan (2001) and followed by Jacob and Lefgren (2005), among others. Appendix A provides a detailed explanation of this analytical correction. While this correction is easy to apply in an OLS model, it becomes considerably more complicated in the dynamic panel context, and we therefore use an instrumental variable strategy for most of our estimators.

3 Data

To demonstrate these issues, 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 3 districts of Punjab: Attock, Faisalabad, and 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. Nevertheless, at the time of the survey, more than 50 percent of the province's population resided in such villages (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, while three refused to cooperate. Sample schools account for over 90 percent 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

---

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

anthropometrics (height and weight) and detailed family characteristics such parental education and wealth, as measured by principal components analysis analysis of 20 assets. When exploring the economic interpretation of persistence, we also use a small subsample of approximately 650 children that can be matched to a detailed household survey that includes, among other things, child and parental time use and educational spending.

For our analysis, we use two subsamples of the data: all children who were tested in all three years (N=8120) and children who were tested *and* given a detailed child survey in all three years (N=4031). 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 exams in English, Urdu (the vernacular), and mathematics. This tests was relatively lengthy (over 40 questions per subject) and was designed to maximize the precision over a range of abilities in each grade. While a fraction of questions changed over the years, the content covered remained consistent, and a significant portion of questions appeared across all years. To avoid the possibility of cheating, the tests were administered directly by our project staff and not by classroom teachers. The tests were scored and equated across years by the authors using Item Response Theory so that the scale has cardinal meaning. 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 attempt to 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. Children were tested in third, fourth, and fifth grades 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.

4 Results

4.1 Cross-sectional and Graphical Results

Before presenting our estimates of learning persistence and the implied private school effect, we provide some rough evidence for a significant private school effect using cross-sectional and graphical evidence. These results don't take advantage of the more sophisticated specifications above but nevertheless provide initial evidence that the zero effect found by assuming perfect persistence and by looking at the trajectories in Figure 1 is implausible.

4.1.1 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 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, even though the R^2 increases substantially. 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 mathematics.

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 English and Urdu. However, even for Urdu, where the female effect is largest, the private school effect is still nearly three times as large. Height, assets, and whether the father (and for Column 3, mother) is educated past elementary school also enter the regression as 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.

4.1.2 Graphical evidence from switching children

Many of the dynamic panel estimators that we explore identify the private school effect 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 after third grade, and switch after fourth grade. For simplicity, we exclude children who switch back and forth between school types.

As the table at the bottom of the figure shows, few children change schools. Only 48 children move from public to private schools in fourth grade, while 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.

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 improve 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 persistence decreases 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.

4.2 OLS and Dynamic Panel Value-Added Estimates

Tables 4 (English), 5 (Urdu), and 6 (mathematics) summarize our main value-added results. All estimates include the full set of controls in the child survey sample, the survey date, round (grade) dummies, and village fixed effects. For brevity, we only report the persistence and private school coefficients.[12] We group the discussion of our results in three domains: estimates of the persistence coefficient, estimates of the private schooling coefficient, and regression diagnostics.

4.2.1 The persistence parameter

We immediately reject the hypothesis of perfect persistence ($\beta = 1$). Across all specifications, except M1 which imposes $\beta = 1$, and all subjects, the estimated persistence coefficient is significantly lower than one, even in the specifications that correct for measurement error only and should be biased upward (M3 and M4). The typical lagged value-added model (M2), which assumes no omitted student heterogeneity and no measurement error, returns estimates between 0.52 and 0.58 for the persistence coefficient. Correcting only for measurement error by instrumenting using the two alternate subjects (M3), or using the analytical correction described by in the appendix (M4), increases the persistence coefficient to between 0.70 and 0.79,

---

[12]As discussed, time-invariant controls drop out of the differenced models. 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.

consistent with significant measurement error attenuation. This estimate, however, remains biased upward by omitted heterogeneity.

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 (M5) or predetermined (M6). In English and Urdu, the persistence parameter falls to between 0.19 and 0.35. The estimates are (statistically) different from models that correct for measurement error only. *In other words, omitted heterogeneity in learning exists, and biases the static estimates upward.* For mathematics, the estimated persistence coefficient is indistinguishable from zero, considerably below all the other estimates. These estimates are higher and somewhat more stable in the systems GMM approach summarized in Panel C (M7, M8).

With the addition of a conditional mean stationarity assumption (Panel D), we can more precisely estimate the persistence 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 doesn't require estimating a system of equations. The persistence coefficient rises substantially to between 0.39 and 0.56. This upward movement is consistent with a violation of the stationarity assumption (the fixed-effect still contributes to achievement growth) but an overall reduction in the omitted heterogeneity bias. Across the various dynamic panel models and subjects, estimates of the persistence parameter vary from 0.2 to 0.55. However the highest dynamic panel estimates come from assuming conditional mean stationary, which is likely too strong in the context of education.

4.2.2   The contribution of private schools

Assuming perfect persistence biases the private school coefficient downward. For English, the estimated private school effect in the restricted model that incorrectly assumes $\beta = 1$ (Panel A, Table 4) is negative and significant. For Urdu and mathematics, the private school coefficient is small and insignificant or marginally significant (Panel A, Tables 5 and 6). By comparison, all the dynamic panel estimates are positive and statistically significant, with the exception of one of the difference GMM estimates, which is too weak to identify the private school effect with any precision.

Panel C (levels and differences GMM) illustrates the benefit of a systems approach. Adding a levels equation (Panel C, Tables 4-6), using the assumption that inputs are constantly correlated or uncorrelated with the omitted effects, reduces the standard errors for the private school coefficient while maintaining the assumption that inputs are predetermined but not strictly exogenous. Under the scenario that private school enrollment is constantly correlated with the omitted effect (M7), the private school coefficient is large: 0.19 to 0.32 standard deviations depending on the subject and statistically significant. This estimate allows for past achievement

shocks to affect enrollment decisions but assumes that switching school type is uncorrelated with unobserved student heterogeneity.

An overarching theme in this analysis is that the persistence parameter influences the estimated private school effect but that it is rarely possible to get enough precision to distinguish estimates based on different exogeneity conditions. This is largely due to the small number of children switching between public and private schools in our sample. In Figure 4, we graph the relationship between both coefficients explicitly. Rather than estimating the persistence 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 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 learning heterogeneity.) As expected given the large baseline differences, the estimated private school effect strongly depends on the assumed persistence rate. Moving from the restricted value-added model ($\beta = 1$) to the pooled cross-section model ($\beta = 0$) increases the estimated effect from negative or insignificant to large and significant. For most of the range of the persistence parameter, the private school effect is positive and significant, but pinning down the precise yearly contribution of private schooling depends on our assumptions about how children learn.

A couple of natural questions are how these estimates compare to the private-public differences reported in the cross-section and why the trajectories in Figure 1 are parallel even though the private school effect is positive. Controlling for observables suggests that after three years, children in private schools are 0.9 (English), 0.5 (Urdu), and 0.45 (mathematics) standard deviations ahead of their public school counterparts. If persistence is 0.4 and the yearly private school effect is 0.3, children's trajectories will become parallel when that achievement gap reaches 0.5 ($= 0.3/1 - 0.4$). This is roughly the gap we find in Urdu and mathematics. Any small disagreement, including the larger gap in English, may be attributable to baseline selection effects. Thus our results can consistently explain the large baseline gap in achievement, the parallel achievement trajectories in public and private schools, and the significant and ongoing positive private school effect.

4.2.3 Regression diagnostics

For many of the GMM estimates, Hansen's J test rejects the overidentifying restrictions implied by the model. This is troubling but not entirely unexpected. As emphasized by Jacob, Lefgren and Sims (2008) in the context of teacher-effect persistence, different instruments may be identifying different local average treatment effects in the education context. 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 more likely identify different parameters than instrumenting using previous year scores. Likewise, instrumenting using alternate lags and differenced achievement and inputs may also identify different effects. This type of heterogeneity is important and suggests that a richer model than a constant coefficient lagged value-added may be warranted.[13] Given the rejection of the overidentifying restrictions in some cases, the next section provides a series of robustness exercises around the estimation of the persistence parameter.

4.3   Robustness Checks

If our estimates are interpreted as forgetting, children lose over half of their achievement in a single year. For some subjects, such as mathematics, this fraction may be even larger. While the estimates reported may appear to be implausibly high, they match recent work on fade-out in value-added models, as well as the rapid fade-out observed in most educational interventions, discussed in Section 5. Furthermore, our view of whether private schools lead to greater learning depends critically on the estimated persistence parameter. Here, we present three robustness exercises in support of our findings.

4.3.1   Experimental evidence of fade-out

Field experiments in education provide an alternate example of achievement decay. Table 7 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. For a number of the interventions, the persistence coefficient is less than 0.10. In two interventions—learning incentives and grade retention—the coefficient is between 0.6 and 0.7. However, these high level of persistence may in part be explained by the specific nature of these interventions.[14] Although the link between fade-out in experimental studies and the persistence parameter is not always exact, all the

---

[13]Another common strategy to address potentially invalid instruments is to slowly reduce the instrument set, testing each subset, until the overidentification test is accepted or the model becomes just identified. We explored this approach but no clear story emerged. One result of note is that dropping the overidentifying inputs typically raises the the persistence coefficient slightly, to roughly 0.25 for math.

[14]In the case of grade retention, there is no real "post treatment" period since children always remain one grade behind after being retained. If one views grade retention as an ongoing multi-period treatment, then lasting effects can be consistent with low persistence. In the case of learning incentives, Kremer, Miguel and Thornton (2003) argue that student incentives increased effort (not just achievement) even after the program ended, leading to ongoing learning. This relates directly to the behavioral interpretation for persistence we discuss in Section 5.

evidence suggests that current learning does not carry over to future learning without loss, and in fact, these losses may be substantial.

4.3.2 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 regression $y_{it} = \alpha + \beta y_{i,t-1} + \eta_{it}$, where we have omitted all potential inputs and corrected only for measurement error bias. Our estimates of this model suggest that the persistence coefficient is at most 0.8 to 0.9—far higher than our highest 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 persistence coefficient is $\mathrm{Cov}(\eta_{it}, y_{i,t-1})/\mathrm{Var}(y_{it-1})$. 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, the persistence coefficient in a lagged value-added model for fourth grade will be biased upward to

$$\frac{\mathrm{Cov}(\eta_{i4}, y_{i3})}{\mathrm{Var}(y_{i3})} = \frac{\rho_X}{2\beta\rho_X - \beta + \beta^2 + 1}. \tag{10}$$

Figure 5 gives a graphical representation of this bias calculation. To read the graph, choose a true persistence 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 persistence coefficient that a lagged value-added specification estimated by OLS would yield. Working with our estimates, if the true persistence effect, β, is 0.4 and inputs are correlated only 0.6 over time, the (incorrectly) estimated β will be 0.9. Given that the vast majority of inputs are fixed, this seems quite reasonable, and perhaps even too low.

4.3.3 Bounding the persistence coefficient using selection on observables

Another way to get at the reasonableness of rapid fade-out is motivated by Altonji, Elder and Taber's (2005) assumption of equal selection on observed and unobserved variables. Altonji et al. suggest that we can learn something about omitted variable bias by exploring how this bias would increase had we omitted the variables we do observe. Modifying Altonji et al.'s idea to address a continuous variable measured with error while letting ε_{it} be the measurement error, equal selection on observables and unobservables can be expressed as

$$\frac{\mathrm{Cov}(\eta_{it}, y_{i,t-1})}{\mathrm{Var}(\eta_{it}) - \mathrm{Var}(\varepsilon_{it})} = \frac{\mathrm{Cov}(\boldsymbol{\alpha}'\mathbf{x}_{it}, y_{i,t-1})}{\mathrm{Var}(\boldsymbol{\alpha}'\mathbf{x}_i)}. \tag{11}$$

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 from Altonji et al.'s original application is that 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.

Table 8 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. Absent controls, the persistence 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 persistence coefficient to 0.74. Thus, just by explaining an additional four percent of the total variation, we reduced the persistence coefficient substantially. 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 lower-bound than an actual estimate, is less than 0.05. Our dynamic panel estimates, therefore, are consistent with the degree of bias we encounter when omitting the variables we do observe.

5   Economic Interpretation

The estimation issues discussed above are largely unrelated to the economic interpretation of the coefficient on lagged test scores. We have used the term persistence to refer to the coefficient in a linear panel data model simply because it is consistent with the education production function framework. But there are various possible explanations for low persistence.

A preliminary concern is that imperfect persistence is a psychometric testing issue. For instance, later test forms may capture fundamentally different latent traits than earlier test forms. To address this concern, we replicated our results using IRT scores based solely on a common set of items that appeared on every test form—our tests had a significant number of overlapping items in each year. Our results were similar using these scores, with the difference GMM persistence estimates in fact dropping slightly. The score equating methods used to create a single cardinal measure of learning therefore does not appear to be the driving force behind the low observed persistence.

There are several other possible mechanical explanations for low persistence. First, artificial ceiling effects can appear like low persistence in models that use bounded scores. To address this concern, we exclusively use unbounded IRT scale scores (see the kernel densities in Figure 2) and our exam is designed to maximize the variation over the entire range of observed abilities. Second, cheating, often driven by high stakes testing, can create artificially low es-

timates of persistence. Jacob and Levitt (2003), for instance, detect teacher cheating in part by looking for poor subsequent performance of students who made rapid gains. In our data, cheating is unlikely both because our test is relatively low stakes and because our project staff administered the exam directly to avoid this possibility. Third, critics of high stakes testing often argue that shallow "teaching to the test" leads to low persistence. This is also an unlikely explanation in our context; our exam is relatively low stakes, is not part of the standard educational infrastructure, and covers only subject matter that all students should know and that Pakistani parents generally demand. Finally, a natural intuition is that even if children's test-measurable skills rapidly decay, children can easily relearn. However, to the degree improved relearning affects subsequent school-based learning, faster relearning should be captured by the persistence parameter we estimate. Whether other attributes not captured by our test persist remains open for further study.

From an economic perspective, low persistence may also be due to behavioral responses. In an optimizing model of household behavior, the lagged coefficient estimate can be interpreted as decay or 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 (or school/teacher) re-optimization following a positive shock or from heterogeneity in initial endowments. Violations of (b) imply that the low coefficient could arise from concavity in the educational production function (i.e., when you know more, it costs more to learn).

Tables 9 and 10 present the results of a preliminary exercise that assesses household and school responses to shocks. 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. This data is extremely rich and includes information on parent's changing perceptions of each child's performance, detailed parental and child time-use data, and detailed expenditure data. 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, expenditure on school, and time spent helping children on homework, being tutored, doing homework, and playing.[15] Table 9 presents the results. Parents' perceptions of their child's performance reacts strongly to gains to achievement, which is likely due to the fact that children's scores were distributed to parents as part of a randomized experiment (see Andrabi et al. (forthcoming) for details on this experiment). However, there is only weak evidence of substitution effects. School expenditures do drop slightly as do the hours spent helping the child on his or her homework. Minutes

---

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

spent playing increases, but tuition also increases. While some of these responses are in the direction of substitution, they are generally not statistically significant. Given the detailed household data we obtained, it suggests that household substitution is unlikely to be a main driver behind low persistence. This may not be particularly surprising given that very low achievement suggests that children may be below parents' desired learning levels.

Table 10 explores a school level response: the possibility that fade-out captures teachers targeting poorly performing students. If teachers target poorly performing children in each classroom, persistence should be lower within schools than between schools. To test this hypothesis, we estimate a basic lagged value-added model with no controls and instrument for lagged achievement using lagged differences in alternate subjects (for simplicity). We estimate this model using average school scores (between school specification) and child scores measured in deviations for the school average (within school specification). If anything, the persistence coefficient is lower for the between school regressions, suggesting that within school targeting is not the primary reason for low persistence.

Although a somewhat uncomfortable position given the lack of consistency with the basic household optimization model, it could just be that children forget what they learned. Psychology and neuroscience provide some compelling evidence for this using laboratory experiments. Psychological research on the "curve of forgetting" dates back to Ebbinghaus's (1885) seminal study on memorization. Rubin and Wenzel (1996) review the laboratory research spawned by this contribution. Semb and Ellis (1994) review classroom studies that test how much students remember after taking a course. Both literatures document the fragility of human memory. Cooper et al. (1996) studies the learning losses that children experience between spring and fall achievement tests. 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.

There is no reason to suspect that all knowledge or all program impacts persists at the same rate. In the education literature, there is also considerable evidence for heterogeneous decay across treatments, and it is possible that the nature of heterogeneity could provide hints on the origins of the low persistence coefficient (Semb and Ellis, 1994). To give one example, MacKenzie and White (1982) report that fade-out for geographical knowledge was much higher for in-class exercises compared to field excursions (or, passive versus active learning.) Similarly, Rothstein (2008) finds heterogeneity in the long-run effects of teachers who produce equal short-run gains; Jacob, Lefgren and Sims (2008) estimate that teacher effects are only a third as persistent as achievement in general.

Using our data, we briefly examine persistence heterogeneity in Table 11. Here, to obtain the most power possible, we estimate the value-added model for specific sub-populations using the "predetermined inputs, uncorrelated effects, and conditionally stationary" based estimator (M10

of Table 4, 5, and 6). 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 government schools, but the difference is not statistically significant. A similar pattern holds for richer families and children with educated parents. These results hint that learning decays faster for faster learners. This is consistent with a forgetting interpretation of persistence where shallow but rapid learners forget at a faster rate than deep but slow learners.

6   Discussion and Conclusion

In the absence of randomized studies, the value-added approach to estimating education production functions has gained momentum as a valid methodology for removing unobserved individual heterogeneity in assessing the contribution of specific programs or in understanding the contribution of school-level factors for learning (e.g. Boardman and Murnane, 1979; Hanushek, 1979; Todd and Wolpin, 2003; Hanushek, 2003; Doran and Izumi, 2004; McCaffrey, 2004; Gordon, Kane and Staiger, 2006). In such models, assumptions about learning persistence and unobserved heterogeneity play central roles. Our results reject both the assumption of perfect persistence required for the restricted value-added model and the no-learning heterogeneity required for the lagged value-added model. Our results for Pakistan should illustrate the danger of incorrectly modeling or estimating education production functions: the restricted value-added model is fundamentally misspecified and can even yield wrong-signed estimates of a program's impact. The original goal of value-added models—to eliminate individual heterogeneity—remains unaccomplished.

Our estimate of persistence is consistent with recent work by Rothstein (2008), Jacob, Lefgren and Sims (2008), and Kane and Staiger (2008), with analytical and empirical estimates of the expected bias under OLS, and with experimental evidence of program fade-out in developing and developed countries. But the economic interpretation still remains an open area of enquiry; while we can argue against several explanations in our context, it is harder to provide compelling evidence in favor of a particular explanation. Our context and test largely rule out cheating and "teaching to the test" explanations of low persistence. By examining achievement measured by a common set of items administered across years, we also rule out imperfect persistence as a psychometric artifact. We find little evidence that low persistence results from substitution by parents and teachers; the behavioral adjustments we are able to measure are unlikely to represent the primary reason achievement gains fade-out. Simple forgetting, consistent with a large body of memory research in psychology, appears to be a likely explanation and hence a core component of education production functions, although more research is needed to provide direct evidence for it.

For non-experimental program evaluation, the conclusion here is somewhat negative. Our

most striking result is the remarkable failure of the restricted value-added approximation of the education production function. 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 offers several alternatives, but these methods 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. For researchers stuck with limited observational data, OLS with the lagged test score included as an explanatory variable and with no correction for measurement error, is superior to the gain score specification, at least for estimating the input effect.

Our results also highlight difficulties with randomized evaluations. The extent of fade-out implies that a short evaluation may yield little information about the cost-effectiveness of a program. Using the one or two year increase from a program gives an upper-bound on the longer term achievement gains. As our estimates suggest, and Table 7 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 three periods of data (a luxury in many evaluations), can theoretically obtain a lower-bound on cost-effectiveness by assuming exponential fade-out. At the same time, the causes of fade-out are equally important: if parents no longer need to hire tutors or buy textbooks (the substitution interpretation of imperfect persistence), a program may be *cost*-effective even if test scores fade out.

Moving forward, empirical estimates of education production functions may benefit from further unpacking persistence. Overall, the agenda pleads for a richer model of education and for empirical techniques for modelling the broader learning process, 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

$$y_{it}^* = \alpha' \mathbf{x}_{it} + \beta y_{i,t-1}^* + v_{it}, \tag{12}$$

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

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

and OLS is inconsistent because $y_{i,t-1}$ is correlated with $\varepsilon_{i,t-1}$.

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

$$\tilde{y}_{i,t-1} \equiv \mathbb{E}^*[y_{i,t-1}^* \mid y_{i,t-1}, \mathbf{x}_{it}] = \lambda' \mathbf{x}_{it} + r_{i,t-1} y_{i,t-1}, \tag{14}$$

where λ and $r_{i,t-1}$ are parameters. To see why this works, add and subtract $\beta \tilde{y}_{i,t-1}$ from (12) to get

$$y_{it} = \mathbf{x}_{it} \alpha + \beta \tilde{y}_{i,t-1} + [\beta(y_{i,t-1} - \tilde{y}_{i,t-1}) + v_{it} + \varepsilon_{it} - \beta \varepsilon_{it-1}] \tag{15}$$
$$= \mathbf{x}_{it} \alpha + \beta \tilde{y}_{i,t-1} + [\beta(y_{i,t-1}^* - \tilde{y}_{i,t-1}) + v_{it} + \varepsilon_{it}]. \tag{16}$$

where the second line follows from $y_{i,t-1}^* = y_{i,t-1} - \varepsilon_{it-1}$. Assuming exogeneity with respect to $v_{it} + \varepsilon_{it}$, OLS is consistent if

$$\mathbb{E}[\mathbf{x}_{it}'(y_{i,t-1}^* - \tilde{y}_{i,t-1})] = 0, \tag{17}$$
$$\mathbb{E}[\tilde{y}_{i,t-1}(y_{i,t-1}^* - \tilde{y}_{i,t-1})] = 0. \tag{18}$$

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 the projection (14).

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 (14) are

$$\mathbb{E}[\mathbf{x}_{it}'(y_{i,t-1}^* - \lambda' \mathbf{x}_{it} - r_{i,t-1} y_{i,t-1})] = 0, \tag{19}$$
$$\mathbb{E}[y_{i,t-1}'(y_{i,t-1}^* - \lambda' \mathbf{x}_{it} - r_{i,t-1} y_{i,t-1})] = 0. \tag{20}$$

Solving first for λ, we have

$$\lambda = \mathbb{E}[\mathbf{x}_{it}' \mathbf{x}_{it}]^{-1} E[\mathbf{x}_{it}'(y_{i,t-1}^* - r_{i,t-1} y_{i,t-1})]. \tag{21}$$

Plugging (21) into (20) and solving for $r_{i,t-1}$ yields

$$r_{i,t-1} \quad = \quad \mathbb{E}[y_{i,t-1}\mathbf{m}_x y_{i,t-1}]^{-1} E[y_{i,t-1}\mathbf{m}_x y^*_{i,t-1}] \tag{22}$$

$$= \quad \mathbb{E}[e^2_{i,t-1}]^{-1} \left(E[e^2_{i,t-1}] - E[\varepsilon^2_{i,t-1}] \right) \tag{23}$$

$$= \quad \frac{\sigma^2_{e_{i,t-1}} - \sigma^2_{\varepsilon_{i,t-1}}}{\sigma^2_{e_{i,t-1}}}, \tag{24}$$

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 (21) to get

$$\lambda \quad = \quad \mathbb{E}[\mathbf{x}'_{it}\mathbf{x}_{it}]^{-1} E[\mathbf{x}'_{it}(y^*_{i,t-1} - r_{i,t-1}y_{i,t-1})] \tag{25}$$

$$= \quad \mathbb{E}[\mathbf{x}'_{it}\mathbf{x}_{it}]^{-1} E[\mathbf{x}'_{it}y_{i,t-1}](1 - r_{i,t-1}). \tag{26}$$

The best predictor is

$$\tilde{y}_{i,t-1} = \mathbb{E}[y_{i,t-1}|\mathbf{x}_{it}](1 - r_{i,t-1}) + r_{i,t-1}y_{i,t-1} \tag{27}$$

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 as 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) .

Table A1 reports persistence coefficients corrected only for measurement error using the instrumental variable (using alternate subjects) and the analytical correction approach. 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 persistence 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 mathematics, and 0.97 for Urdu. However, the overidentifying restriction is rejected at the one 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 A1 shows estimates based on our analytical correction around 0.9. Of course, all of these estimates remain biased upward by learning heterogeneity.

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.

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).

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.

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. "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* .

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

Deming, David. 2008. "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.

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* .

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.

Harris, D. and T.R. Sass. 2006. "Value-Added Models and the Measurement of Teacher Quality." *Unpublished manuscript* .

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* .

Jacob, Brian, Lars John Lefgren and David Sims. 2008. "The Persistence of Teacher-Induced Learning Gains." *NBER Working Paper* .

Jacob, Brian and S.D. Levitt. 2003. "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *The Quarterly Journal of Economics* 118(3):843–877.

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.

Kane, T.J. and D.O. Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." *Unpublished. Cambridge, MA: Harvard University* .

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

Rothstein, Jesse. 2008. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *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.

Schweinhart, L. J., J. Montie, Z. Xiang, W. S. Barnett, C. R. Belfield and M. Nores. 2005. *Lifetime effects: The High/Scope Perry Preschool study through age 40.* Ypsilanti, MI: High/Scope Press.

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

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.

FIGURE 1. EVOLUTION OF TEST SCORES IN PUBLIC AND PRIVATE SCHOOLS
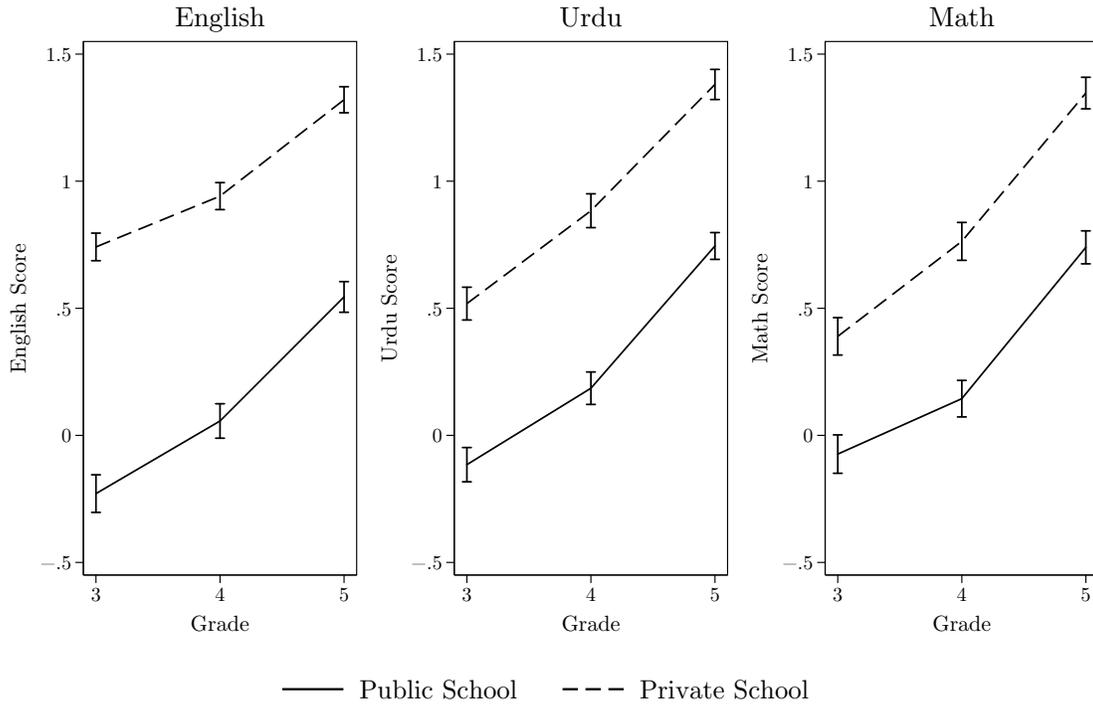
Notes: Vertical bars represent 95% confidence intervals around the group means, allowing for arbitrary clustering within schools. Tests scores are IRT based scale scores normalized to have mean zero and standard deviation one for the full sample of children in third grade. Children who were tested in 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

English

Urdu

Math
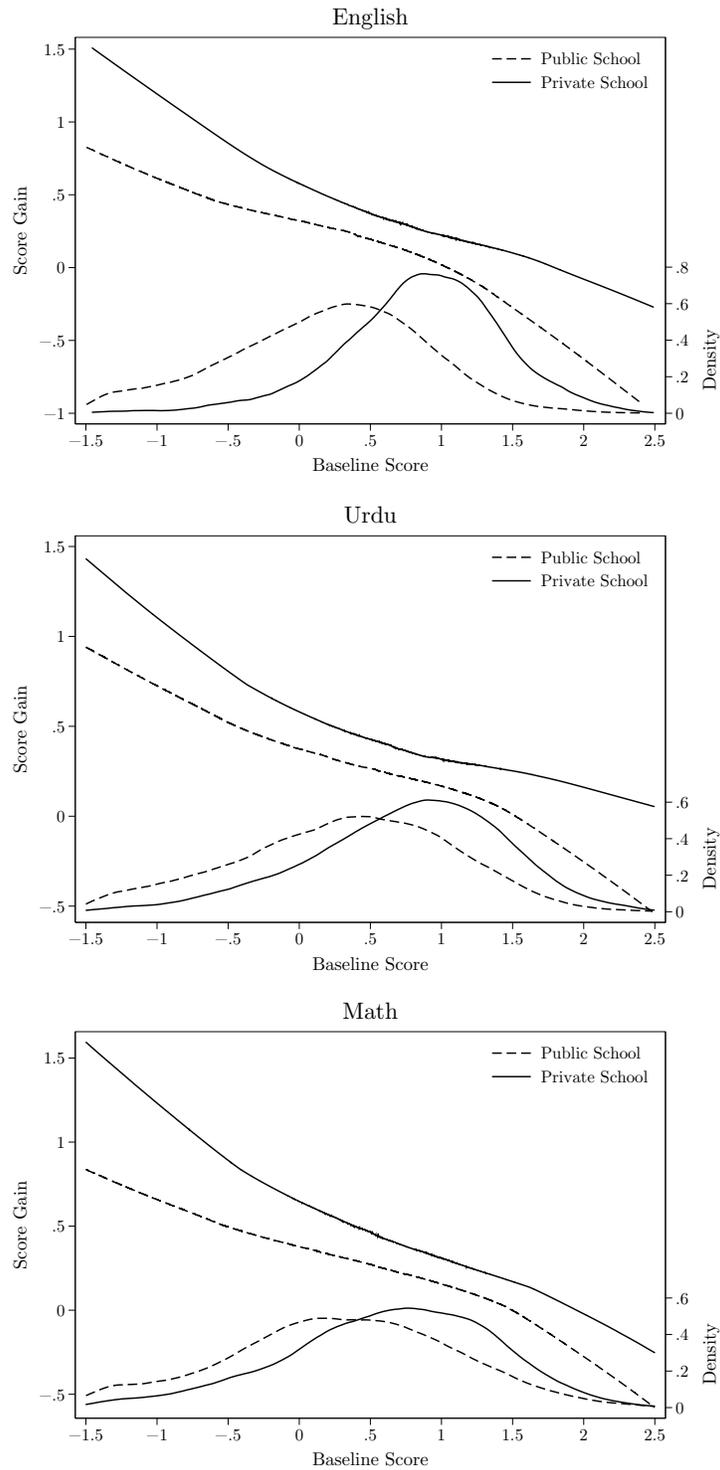
Notes: The yearly score gain vs. baseline score lines for public and private schools are smoothed using a locally-weighted regression estimator. The sample includes children tested in all periods with baseline score between -1.5 and 2.5 (to avoid noise near tails). The kernel density for children's scores 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			
M1. No depreciation β=1 (OLS)	n/a	n/a	Assumes perfect persistence and no or uncorrelated heterogeneity.
M2. No effects, no measurement error (OLS)	n/a	n/a	Assumes no measurement error and no effects.
M3. No effects (2SLS/IV correction)	n/a	Alternate subjects	Assumes no effects and uncorrelated measurement errors across subjects
M4. No effects (HEIV/ Analytical correction)	n/a	n/a	Assumes no effects and analytical correction is valid.
Panel B: Difference GMM			
M5. Strictly exogenous inputs	Inputs: 1...T Score: 1...t-2	n/a	Assumes no feedback effects.
M6. 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			
M7. Predetermined inputs, constantly correlated effects	Inputs: 1...t-1 Score: 1...t-2	ΔInputs: 1..t	Assumes effects have constant correlation with inputs.
M8. 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)			
M9. Predetermined inputs, constantly correlated effects, conditional stationarity	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.
M10. Predetermined inputs, uncorrelated effects, conditional stationarity	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
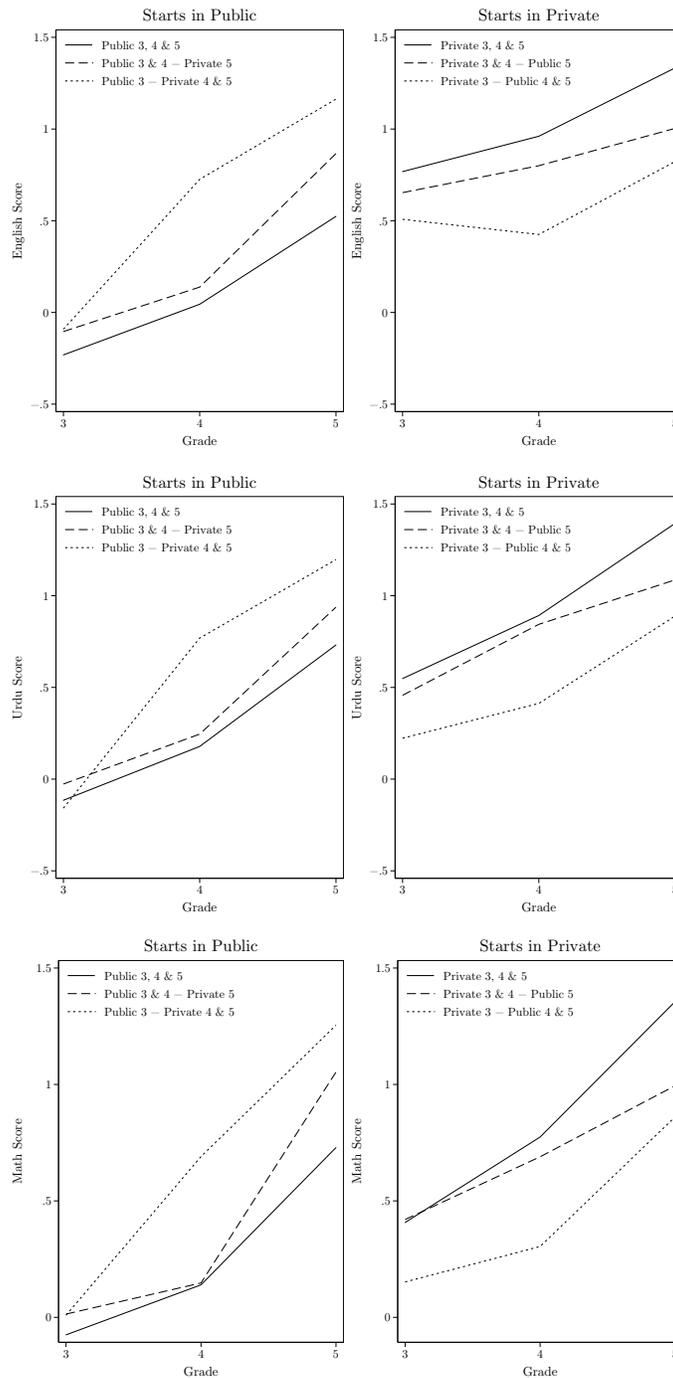
	Public 3, 4, & 5	Public 3, 4 − Private 5	Public 3 − Private 4 & 5	Private 3, 4 & 5	Private 3 & 4 − Public 5	Private 3 − Public 4, 5
N	5688	40	48	2007	160	167

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. VALUE-ADDED MODEL ESTIMATES OF PERSISTENCE AND PRIVATE SCHOOL COEFFICIENT (ENGLISH)

Estimator's Key Assumption	Persistence Coefficient	Private School Coefficient	Hansen's J χ² (p-value)	df
Panel A: Static Estimates				
M1. No depreciation β=1 (OLS)	1.00	-0.08 (0.02)		
M2. No effects, no measurement error (OLS)	0.52 (0.02)	0.31 (0.02)		
M3. No effects (2SLS/IV Correction)	0.70 (0.02)	0.16 (0.02)	4.69 (0.03)	1
M4. No effects (HEIV/Analytical Correction)	0.74 (0.02)	0.21 (0.02)		
Panel B: Difference GMM				
M5. Strictly exogenous inputs	0.19 (0.10)	0.25 (0.07)	25.44 (0.02)	13
M6. Predetermined inputs	0.19 (0.10)	1.15 (0.39)	16.82 (0.02)	7
Panel C: Levels and Difference SGMM				
M7. Predetermined inputs, constantly correlated effects	0.36 (0.07)	0.21 (0.06)	45.50 (0.00)	23
M8. Predetermined inputs, uncorrelated effects	0.53 (0.05)	0.32 (0.04)	79.08 (0.00)	29
Panel D: Levels Only GMM				
M9. Predetermined inputs, constantly correlated effects, conditional stationarity	0.40 (0.05)	0.29 (0.07)	24.74 (0.02)	12
M10. Predetermined inputs, uncorrelated effects, conditional stationarity	0.39 (0.05)	0.24 (0.04)	23.43 (0.02)	11

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 and Table 1 for details on instruments and assumptions.

TABLE 5. VALUE-ADDED MODEL ESTIMATES OF PERSISTENCE AND PRIVATE SCHOOL COEFFICIENT (URDU)

Estimator's Key Assumption	Persistence Coefficient	Private School Coefficient	Hansen's J χ^2 (p-value)	df
Panel A: Static Estimates				
M1. No depreciation $\beta=1$ (OLS)	1.00	0.01 (0.02)		
M2. No effects, no measurement error (OLS)	0.58 (0.01)	0.26 (0.02)		
M3. No effects (2SLS/IV Correction)	0.73 (0.02)	0.17 (0.02)	3.67 (0.06)	1
M4. No effects (HEIV/Analytical Correction)	0.79 (0.02)	0.20 (0.02)		
Panel B: Difference GMM				
M5. Strictly exogenous inputs	0.21 (0.09)	0.29 (0.07)	49.50 (0.00)	13
M6. Predetermined inputs	0.35 (0.11)	0.90 (0.48)	18.90 (0.01)	7
Panel C: Levels and Difference SGMM				
M7. Predetermined inputs, constantly correlated effects	0.26 (0.08)	0.22 (0.06)	66.58 (0.00)	23
M8. Predetermined inputs, uncorrelated effects	0.51 (0.06)	0.30 (0.04)	81.89 (0.00)	29
Panel D: Levels Only GMM				
M9. Predetermined inputs, constantly correlated effects, conditional stationarity	0.55 (0.05)	0.31 (0.07)	13.49 (0.33)	12
M10. Predetermined inputs, uncorrelated effects, conditional stationarity	0.56 (0.05)	0.27 (0.03)	13.30 (0.27)	11

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 and Table 1 for details on instruments and assumptions.

TABLE 6. VALUE-ADDED MODEL ESTIMATES OF PERSISTENCE AND PRIVATE SCHOOL COEFFICIENT (MATH)

| Estimator's Key Assumption	Persistence Coefficient	Private School Coefficient	Hansen's J χ^2 (p-value)	df
Panel A: Static Estimates				
M1. No depreciation $\beta=1$ (OLS)	1.00	0.05 (0.02)		
M2. No effects, no measurement error (OLS)	0.57 (0.02)	0.27 (0.03)		
M3. No effects (2SLS/IV Correction)	0.76 (0.02)	0.17 (0.03)	0.02 (0.89)	1
M4. No effects (HEIV/Analytical Correction)	0.75 (0.02)	0.23 (0.03)		
Panel B: Difference GMM				
M5. Strictly exogenous inputs	-0.00 (0.09)	0.26 (0.08)	33.97 (0.00)	13
M6. Predetermined inputs	0.12 (0.12)	0.46 (0.50)	12.06 (0.10)	7
Panel C: Levels and Difference SGMM				
M7. Predetermined inputs, constantly correlated effects	0.12 (0.10)	0.19 (0.08)	57.63 (0.00)	23
M8. Predetermined inputs, uncorrelated effects	0.51 (0.08)	0.30 (0.05)	82.19 (0.00)	29
Panel D: Levels Only GMM				
M9. Predetermined inputs, constantly correlated effects, conditional stationarity	0.51 (0.06)	0.30 (0.07)	29.45 (0.00)	12
M10. Predetermined inputs, uncorrelated effects, conditional stationarity	0.53 (0.06)	0.27 (0.04)	28.36 (0.00)	11

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 and Table 1 for details on instruments and assumptions.

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

English

Urdu

Math
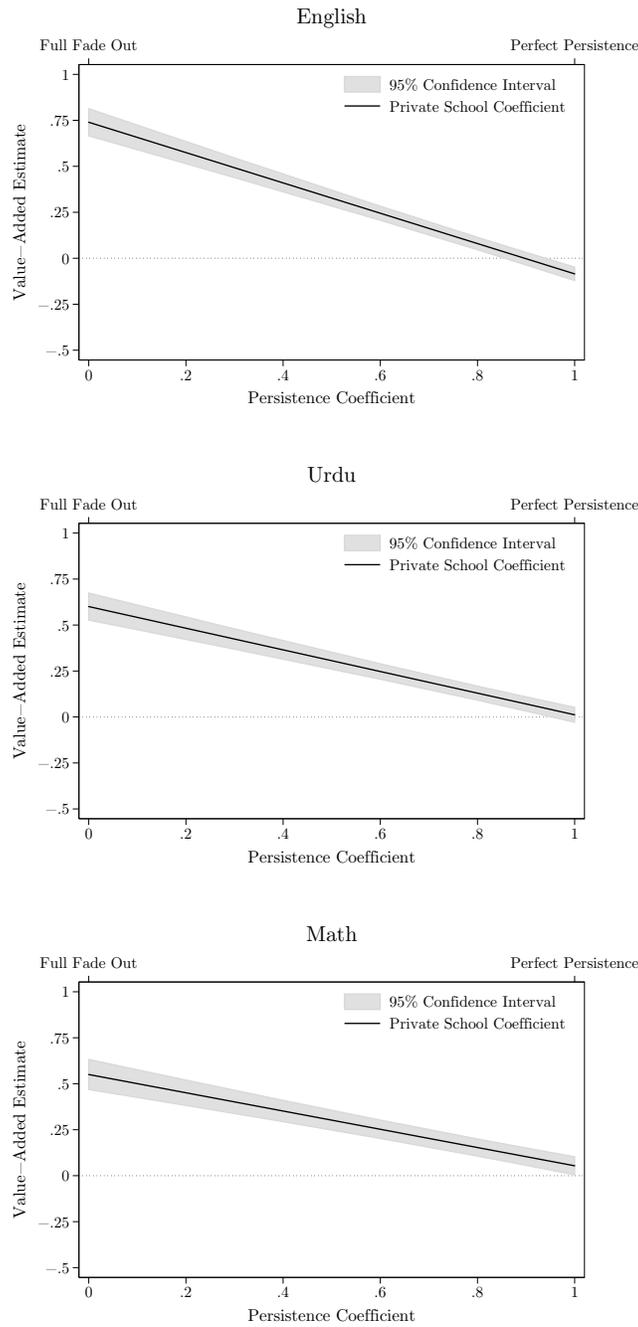
Notes: These graphs show the estimated value-added effect of private schools depending on the assumed persistence coefficient of lagged achievement. The restricted value-added model, for example, assumes the persistence coefficient equals one—no fade-out. The estimated value-added pooled for third to 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.

TABLE 7. EXPERIMENTAL ESTIMATES OF PROGRAM FADE OUT

Program	Subject	Immediate Treatment Effect	Extended Treatment Effect	Implied Persistence Coefficient	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	Math	0.136	0.095	0.70	Jacob and Lefgren
Grade Retention	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.

FIGURE 5. TRUE AND ESTIMATED PERSISTENCE IN A LAGGED VALUE-ADDED MODEL WITH
SERIALLY CORRELATED OMITTED INPUTS
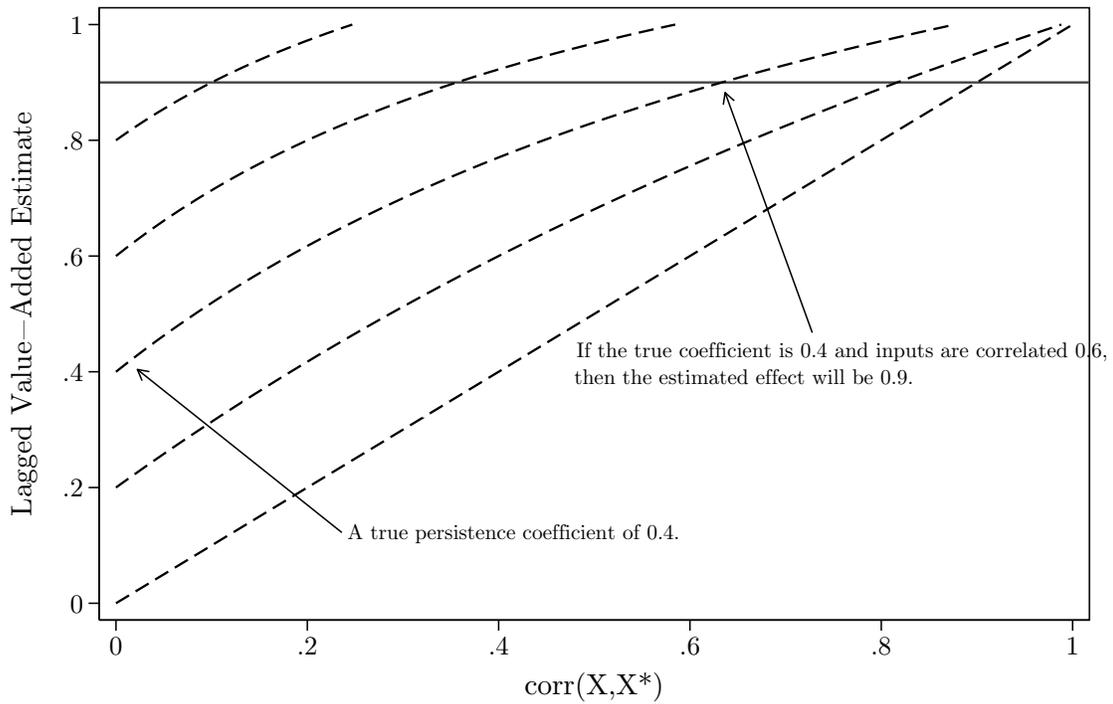
Notes: In the lagged value-added model, the persistence coefficient is biased upward by the correlation between omitted contemporaneous inputs and past inputs that are captured in the lagged test score. Assuming constant correlation between any two years of inputs X and X*, the bias can be calculated analytically (see text). The graph above gives the implied bias for fourth grade. The dashed lines represent the true persistence 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 persistence rate (dashed lines) and on the assumed correlation of inputs over time (x-axis). For example, a (biased) estimated persistence coefficient of 0.9 may result from a true persistence coefficient of 0.4 and correlation between inputs around 0.6. These calculations that assume achievement is measured perfectly and that all inputs are omitted (i.e. unobserved) in the regression.

Table 8. Persistence 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 unobservables, 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 9. HOUSEHOLD RESPONSES TO PERFORMANCE SHOCKS

| Household Changes (Grade 4 to 5) | Test Score Residual (Grade 4) | | | |
	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 (tutoring)	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 10. PERSISTENCE COEFFICIENT USING WITHIN AND BETWEEN SCHOOL VARIATION ONLY

Subject	Variation	Persistence 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: Persistence 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 11. Persistence Coefficient Heterogeneity Across School, Child and Household Characteristics

	Persistence Coefficient		
Within category:	**English**	**Urdu**	**Math**
Private School	0.35 (0.09)	0.45 (0.09)	0.31 (0.11)
Public School	0.41 (0.05)	0.56 (0.06)	0.55 (0.07)
Female	0.48 (0.08)	0.59 (0.08)	0.53 (0.08)
Male	0.37 (0.06)	0.51 (0.06)	0.50 (0.08)
Richer Family	0.34 (0.09)	0.57 (0.10)	0.33 (0.11)
Poorer Family	0.40 (0.09)	0.52 (0.08)	0.55 (0.09)
Mother educated past primary	0.43 (0.12)	0.49 (0.14)	0.44 (0.11)
Mother not education past primary	0.40 (0.05)	0.57 (0.05)	0.52 (0.07)
Father educated past primary	0.49 (0.07)	0.56 (0.08)	0.48 (0.08)
Father not educated beyond primary	0.34 (0.07)	0.54 (0.06)	0.54 (0.08)

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.

TABLE A1. 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.