The Persistence of Teacher-Induced Learning Gains

Brian A. Jacob
Harvard University and NBER

Lars Lefgren
Brigham Young University

David Sims
Brigham Young University

April 2008

Abstract

Most contemporary economics of education literature specifies reduced from educational production functions as approximations of the learning process. These models likely obscure differences in the decay paths of different shocks to learning. For example, recent literature on teacher quality measured as test score value added implies that policies designed to raise teacher value added may greatly improve student learning. In this paper we develop a simple statistical model in which learning consists of both transitory and permanent components and demonstrate that an instrumental variables estimator can be used to recover the fraction of learning added by teachers that persists over time. We compare this estimate with two benchmarks for the more general persistence of student gains in knowledge. We find that the vast majority of the contemporary test score effect attributed to teacher value-added is transitory. This suggests that the teacher value-added literature overstates the effect of teachers on long-run learning and, therefore, the ability of policies that target teacher value-added to change ultimate student outcomes. This method is easily generalized to compare the persistence of other interventions.

1. Introduction

Educational interventions are often narrowly targeted and temporary, such as class size reductions in kindergarten or summer school in selected elementary grades. Because of financial, political and logistical constraints, evaluations of such programs often focus exclusively on the short-run impacts of the intervention. Insofar as the treatment effects are immediate and permanent, short-term evaluations will provide a good indication of the long-run impacts of the intervention. However, prior research such as the Currie and Thomas work on Head Start (1995) suggests that the positive effects of educational interventions may fade out over time. Failure to account for this fade out can dramatically change the assessment of the program impact and/or cost effectiveness.

In this paper, we develop a simple statistical framework to empirically assess the persistence of treatment effects in education in the context of teacher value-added and show how the approach can be generalized to apply to other educational interventions. Measuring the persistence in a teacher's ability to improve student test scores is an important application due to the recent focus of education researchers, practitioners and policymakers on using value-added measures of teacher performance for policy design and accountability.

Using an administrative education data set we construct measures of teacher value added and estimate the persistence of value added effects on student test scores. We find that gains in math and reading test scores due to teacher ability quickly erode. In most cases, our point estimates suggest a one-year persistence of about one-fifth and rule out a one-year persistence rate higher than one-third. Our results are robust to a number of specification checks and suggest that this depreciation applies to almost all student groups.

Comparisons with the general persistence of student ability suggest teacher influence is only a third as persistent as learning innovations in general. Further estimates suggest that about one-eighth of the original student gains from a high value added teacher persist over two years. This evidence suggests that even if value added models of teacher quality are econometrically modified to work well in measuring one period gains, the results will still be misleading in policy evaluation if that single period measure is taken as an indication of the long run increase in knowledge. While there is little evidence indicating how costly it might be for teachers to improve their individual value added, there may be few long term student academic benefits realized from reward or compensation schemes based on these measures.

The remainder of the paper proceeds as follows. Section 2 discusses the motivation for examining the persistence of teacher value added, section 3 introduces the statistical model of student learning, section 4 outlines the data, section 5 presents the results and a short discussion, while section 6 concludes.

2. Background

A. Teacher value added

Despite a widespread belief among education practitioners and the public about the important role of teachers in promoting student achievement, an initial generation of research widely confirmed the Coleman et al. (1966) report's conclusion that there was little association between measurable teacher characteristics and student achievement.

Indeed, with the exception of a notable improvement in teacher performance associated with the first year or two of experience (Hanushek 1997) researchers were left to justify why schools and teachers "don't seem to matter." (Goldhaber and Brewer 1997).

More recently, the growing availability of longitudinal, student achievement data linked to teachers has allowed researchers to calculate sophisticated value-added models that attempt to isolate an individual teacher's contribution to student learning. These studies consistently find substantial variation in teacher effectiveness. For example, the findings of Rockoff (2004) and Rivkin, Hanushek and Kain (2005) both suggest a one standard deviation increase in teacher quality improves student math scores at least 0.1-0.15 standard deviations. Aaronson, Barrow and Sander (2007) find similar results using high school data. In comparison, this suggests that a one standard deviation increase in teacher quality, as measured by value-added, improves contemporary student test scores as much as a 4-5 student decrease in class size.

The results of these studies have led many researchers and policymakers to promote policies to increase the effectiveness of classroom teachers, such as compensation policy and tenure reviews. (Doran and Izumi (2004), McCaffrey (2004)). Indeed, the inherent optimism of this literature is captured by an oft-cited statistic that matching a student with a stream of good teachers (one standard deviation above the average teacher) for 5 years in a row would be enough to complete eliminate the achievement gap between poor and non-poor students (Rivkin, Hanushek and Kain 2005). Given the poor record of single year test scores (Kane and Staiger 2002) or even principal evaluations (Jacob and Lefgren 2005a) in differentiating among certain regions of the teacher quality distribution, the increasing use of value added measures seems likely wherever the data requirements can be met.

However, this research measuring the specific contribution of teachers to student achievement is only one strand of a broader literature utilizing value added estimation. The cumulative nature of knowledge suggests that a current test score is in fact a function of student characteristics combined with the characteristics and policy innovations of all schools and classrooms the student has been in to date. This creates a serious risk that unmeasured past factors will bias estimates of any non-experimental intervention. The most common response since Boardman and Murnane (1979) has been the value added approach whereby the researcher accounts for the past achievement of a student, either by using a within student model differenced across time, or by controlling for a lagged test score measure. This type of specification was widely believed to substantially reduce the chance of bias due to historical omitted variables (Hanushek 2003).

A number of recent studies (Andrabi et. al.(2008), McCaffrey et. al. (2004), Rothstein (2007), Todd and Wolpin 2003, 2006) have highlighted the strong assumptions of the value added teacher model and suggested they are unlikely to hold in observational settings. The most important of these assumptions in our present context is that the assignment of students to teachers is random. Indeed given random assignment of students to teachers, many of the uncertainties regarding precise functional form become less important. If students are not assigned randomly to teachers, positive outcomes attributed to a given teacher may simply result from teaching better students. In particular, Rothstein (2007) raises disturbing questions about the validity of current teacher value added measurements, showing that the current performance of students can be predicted by the value added of their future teachers.

However, in a recent attempt to validate observationally derived value added methods with experimental data, Kane and Staiger (2008) were unable to reject the hypothesis that the observational estimates were unbiased predictions of student achievement in many specifications. Indeed, one common result seems to be that models which control for lagged test scores, such as our model, tend to perform better than gains models. While we are still concerned about the possible consistency of our value added estimates in the presence of possible non-random matching of students to teachers, we will argue that at a minimum our estimates still present a useful upper bound to the true persistence of teacher effects on student achievement.

B. Persistence

As Todd and Wolpin (2003) note, many value added studies implicitly make a strong assumption by restricting the rate of decay of an input induced achievement gain to either zero or a constant. More importantly, the model as commonly specified does not recognize that the rate of decay might depend on the nature of the input. This is important since previous research on the long term impacts of educational interventions suggest decay may vary widely by type of program. For example, long term follow up studies of some programs the Tennessee class size experiment (Nye, Hedges and Konstantopoulos 1999; Krueger and Whitmore 2001) or the Perry preschool project (Barnett 1985) suggest that both had enduring measurable effects, in the later case decades later. On the other hand, evaluations of other similar programs such as head start (Currie and Thomas 1995) or grade retention for sixth graders (Jacob and Lefgren 2004) find no measurable effects on students a few years later.

Furthermore, these studies provide no systematic way to think about comparing persistence across programs, or to test hypotheses about persistence. Most commonly, persistence is inferred as the informal ratio of coefficients from separate regressions. This paper presents a more systematic alternative.

Much of the research on teacher value added also fails to consider the importance of persistence either as an absolute policy parameter or relative to other programs. Counterfactual comparisons, such as the Rivkin, Hanushek and Kain (2005) five good teachers scenario explicitly assume perfect persistence of student gains due to teacher quality and treat test score increases from this source as equivalent to those due to increased parental investment or innate student ability.

Despite the importance of measuring the persistence of teacher value added there has been relatively little work to date on the subject. The first paper to explicitly consider the issue was a study by McCaffery et al (2004) Although their results imply a rate of decay similar to our findings, their small sample results in standard errors that are in general equal to or greater then the estimated coefficients, allowing a wide range of possible persistence values. The only other study to explicitly treat this topic of which we are aware is an unpublished working paper by Kane and Staiger (2008) where the authors use coefficient ratios from OLS regressions to examine persistence.[1]

[1] Although it is not the point of his paper, Rothstein (2007) mentions the importance of measuring fade out and presents evidence of a two-year fade out of approximately one-half in “classroom effects.” However, he is unable to separate the teacher components from other classroom achievement shocks with data from a single cohort of students.

3. A Statistical Model

This section outlines a simple model of student learning that incorporates permanent as well as transitory learning gains. Our goal is to explicitly illustrate how learning in one period is related to knowledge in subsequent periods. Using this model, we demonstrate how the parameter of interest, the persistence of a particular measurable education input, can be recovered via instrumental variables as a particular local average treatment effect (Imbens and Angrist 1994). We initially motivate this strategy in the context of teacher quality but then generalize the model to take into account arbitrary educational interventions.

In order to control for past student experiences, education researchers often employ empirical strategies that regress current achievement on lagged achievement, namely

(1) $Y_t = \beta Y_{t-1} + \varepsilon_t$,

with the common result that the OLS estimate of beta is less than one. This result is typically given one of two interpretations. One explanation is that the lagged achievement score is measured with error due to factors such as guessing or test conditions. A second explanation involves the depreciation or decay of knowledge over time, which is typically assumed to be constant.

In order to explore the persistence of knowledge, it is useful to more carefully articulate the learning process underlying these test scores. To begin, suppose that true knowledge in any period is a linear combination of what "long-term" and "short-term" knowledge, which we label with the subscripts l and s. With a t subscript to identify time period this leads to the following representation:

(2). $Y_t = y_{l,t} + y_{s,t}$.

As the name suggests, long-term knowledge remains with an individual for multiple periods, but is allowed to decay over time. Specifically, we assume that it evolves according to the following process:

(3) $y_{l,t} = \delta y_{l,t-1} + \theta_{l,t} + \eta_{l,t}$,

where δ indicates the rate of decay and is assumed to be less than one in order to make y_l stationary.[2] The second term, $\theta_{l,t}$, represents a teacher's contribution to long -term knowledge in period t. The final term, $\eta_{l,t}$, represents idiosyncratic factors affecting long-term knowledge.

In contrast, short-term knowledge reflects skills and information a student has in one period that decay entirely by the next period.[3] Short-run knowledge evolves according to the following process:

(4) $y_{s,t} = \theta_{s,t} + \eta_{s,t}$,

which mirrors equation (3) above when delta, the persistence of knowledge, is zero. Here, the term $\theta_{s,t}$ represents a teacher's contribution to the stock of short-term knowledge and $\eta_{s,t}$ captures other factors that affect short-term performance. The same factors that affect the stock of long-term knowledge could also impact the amount of short-term knowledge. For example, a teacher may help students to internalize some concepts, while only briefly presenting others immediately prior to an exam. The former concepts

[2] This assumption can be relaxed if we restrict our attention to time-series processes of finite duration. In such a case, the variance of $y_{l,t}$ would tend to increase over time.

[3] The same piece of information may be included as a function of either long-term or short-term knowledge. For example, a math algorithm used repeatedly over the course of a school year may enter long term knowledge. Conversely, the same math algorithm, briefly shown immediately prior to the administration of an exam, could be considered short term knowledge.

likely form part of long-term knowledge while the latter would be quickly forgotten. Thus it is likely a given teacher affects both long and short-term knowledge, though perhaps to different degrees.

While they may be conceptually different, observed variation in knowledge due to measurement error and observed variation due to the presence of short run (perfectly depreciable) knowledge are observationally equivalent in this model. For example, a teacher cheating on behalf of students would appear to increase short-term knowledge. Alternatively, a teacher who effectively helps students internalize a concept which is tested in only a single year would appear to increase short-term as opposed to long term knowledge.[4] The observational equivalence of measurement error and short run knowledge is a consequence of limitations in the ability to measure achievement, even though it does not directly affect the conclusions we draw in this paper.

In most empirical contexts, the researcher only observes the total of long and short run knowledge, $Y_t = y_{l,t} + y_{s,t}$. This is consistent with observing a single test score which captures some combination of long and short run knowledge. For simplicity we initially assume that $\theta_{l,t}$, $\eta_{l,t}$, $\theta_{s,t}$, and $\eta_{s,t}$ are independently and identically distributed, although we will relax this assumption later.[5] It is then straightforward to show that when considering this composite test score in the typical "value added" regression model given by equation (1), the OLS estimate of β converges to:

[4] This presupposes that understanding the concept does not facilitate the learning of a more advanced concept which is subsequently tested. For example, even though simple addition may only be tested in early grades, mastery of such material would facilitate the learning of more advanced methods.

[5] Note that both the process for long run and short run knowledge accumulation are stationary implying children have no upward learning trajectory. This is clearly unrealistic. The processes, however, can be reinterpreted as deviations from an upward trend.

(5) $$p\lim\left(\hat{\beta}_{OLS}\right)=\delta\frac{\sigma_{y_l}^2}{\sigma_{y_l}^2+\sigma_{y_s}^2}=\delta\frac{\sigma_{\theta_l}^2+\sigma_{\eta_l}^2}{\left(1-\delta\right)\left(\sigma_{\theta_s}^2+\sigma_{\eta_s}^2\right)+\sigma_{\theta_l}^2+\sigma_{\eta_l}^2}.$$

Thus, OLS identifies the persistence of long run knowledge multiplied by the fraction of variance in total knowledge attributable to long run knowledge. Perhaps a more intuitive way to describe this is the OLS coefficient measures the average persistence of observed knowledge. The formula above also illustrates the standard attenuation bias result if we reinterpret short-term knowledge as measurement error.

This model also allows us to leverage different identification strategies to recover alternative parameters of the data generating process. Suppose, for example, that we estimate equation (3) using instrumental variables with a first stage relationship given by:

(6) $$Y_{t-1}=\pi Y_{t-2}+v_t.$$

We will refer to the estimate of β from this identification strategy as $\hat{\beta}_{LR}$, where the subscript is an abbreviation for long-run. It is again straightforward to show that this estimate converges to:

(7) $$p\lim\left(\hat{\beta}_{LR}\right)=\delta,$$

which is simply the persistence of long-run knowledge.

Most importantly, consider what happens if we instrument lagged knowledge, Y_{t-1} with the lagged teacher's contribution (value-added) to total lagged knowledge, $\Theta_{t-1}=\theta_{l,t-1}+\theta_{s,t-1}$. The first stage is given by:

(8) $$Y_{t-1}=\pi\Theta_{t-1}+v_t.$$

In this case, the second stage estimate, which we refer to as $\hat{\beta}_{VA}$ converges to:

(9) $$p\lim\left(\hat{\beta}_{VA}\right)=\delta\frac{\sigma_{\theta_l}^2}{\sigma_{\theta_l}^2+\sigma_{\theta_s}^2}.$$

The interpretation of this estimator becomes simpler if we think about the dual role of teacher quality in our model. Observed teacher value added varies for two reasons—the teacher's contribution to long-term knowledge and the contribution to short-term knowledge. $\hat{\beta}_{VA}$ measures the fraction of variation in teacher quality attributable to long-term knowledge creation.

Fundamentally, the differences in persistence identified by the three estimation procedures above are a consequence of different sources of identifying variation. For example, estimation of $\hat{\beta}_{OLS}$ generates a persistence measure that reflects all sources of variation in knowledge, from barking dogs to parental attributes to policy initiatives. On the other hand, an instrumental variables strategy isolates variation in past test scores due to a particular factor or intervention. Consequently, the estimated persistence of achievement gains can vary depending on the chosen instrument, as each identifies a different local average treatment effect. In our example $\hat{\beta}_{VA}$ measures the persistence in test scores due to variation in teacher value added in isolation from other sources of test score variation.

This suggests a straightforward generalization: to identify the coefficient on lagged test score using an instrumental variable strategy, one can use any factor that is orthogonal to ε_t as an instrument for y_{it-1} in identifying β. Thus, for *any* educational intervention for which assignment is uncorrelated to the residual, one can recover the persistence of treatment-induced learning gains by instrumenting lagged performance with lagged treatment assignment. Within the framework above, suppose that

$\theta_{lt} = \gamma_l treat_t$ and $\theta_{st} = \gamma_s treat_t$, where γ_l and γ_s reflect the treatment's impact on long and short-term knowledge respectively. In this case, instrumenting lagged observed knowledge with lagged treatment assignment yields an estimator which converges to the following:

$$(10) \qquad p\lim\left(\hat{\beta}_{TREAT}\right) = \delta \frac{\gamma_l}{\gamma_l + \gamma_s}.$$

The estimator reflects the persistence of long-term knowledge multiplied by the fraction of the treatment related test score increase attributable to gains in long-term knowledge.

Beyond the assurance that we are recovering the parameter of interest, our approach has a number of advantages over the informal examination of coefficient ratios often used to think about persistence. It allows us to examine the persistence of policy induced learning shocks relative to intuitive baselines, transformative learning or a change in ability and a "business as usual" index of educational persistence. It also provides a straightforward way to conduct inference on persistence measures. Furthermore, the methodology can be applied to compare persistence among policy choices.

Returning to our examination of the persistence of teacher-induced learning gains, we relax some assumptions regarding our data generating process to highlight alternative coefficient interpretations and threats to identification. First, consider a setting in which an intervention's effect on long and short-term knowledge are not independent. In that case $\hat{\beta}_{VA}$ converges to:

$$(11) \qquad p\lim\left(\hat{\beta}_{VA}\right) = \delta \frac{\sigma^2_{\theta_l} + \operatorname{cov}\left(\theta_l, \theta_s\right)}{\sigma^2_{\theta_l} + \sigma^2_{\theta_s} + 2\operatorname{cov}\left(\theta_l, \theta_s\right)} = \delta \frac{\operatorname{cov}\left(\theta_l, \Theta\right)}{\sigma^2_{\Theta}}.$$

While δ maintains the same interpretation, the remainder of the expression is equivalent to the coefficient from a bivariate regression of θ_l on Θ. In other words, it captures the rate at which a teacher's impact on long term knowledge increases with total measured knowledge.

Another interesting consequence of relaxing this independence assumption is that β_{VA} need not be positive. In fact, if $\operatorname{cov}(\theta_l, \theta_s) < -\sigma^2_{\theta_l}$, β_{VA} will be negative. This can only be true if $\sigma^2_{\theta_l} < \sigma^2_{\theta_s}$. This would happen if observed value added captured primarily a teacher's ability to induce short term gains in achievement and this is negatively correlated to a teacher's ability to raise long term achievement. Although this is an extreme case, it is clearly possible and serves to highlight the importance of understanding the long run impacts of teacher value-added.[6]

Although, relaxing the independence assumption does not violate any of the restrictions for satisfactory instrumental variables identification, β_{VA} can no longer be interpreted as a true persistence measure. Instead, it identifies the extent to which teacher-induced achievement gains predict subsequent achievement.

However, there are some threats to identification that we initially ruled out by assumption. For example, suppose that $\operatorname{cov}(\theta_{l,t}, \eta_{l,t}) \neq 0$, as would occur if children with unobserved high learning ability are systematically allocated to the best teachers. The opposite could occur if principals assign the best teachers to children with the lowest learning potential. In either case the effect on our estimate depends on the sign of the covariance, since:

[6] Jacob and Levitt (2003) find evidence of teacher cheating in Chicago. This cheating, which led to large observed performance increases, was correlated to poor actual performance in the classroom.

(12) $$p\lim\left(\hat{\beta}_{VA}\right) = \delta \frac{\sigma_{\theta_l}^2 + \operatorname{cov}\left(\theta_l, \eta_l\right)}{\sigma_{\theta_l}^2 + \sigma_{\theta_s}^2}.$$

If students with the best idiosyncratic learning shocks are matched with high quality teachers, the estimated degree of persistence will be biased upwards. In the context of standard instrumental variables estimation, lagged teacher quality fails to satisfy the necessary exclusion restriction because it affects later achievement through its correlation with unobserved educational inputs. To address this concern, we show the sensitivity of our persistence measures to the inclusion of student-level covariates, which would be captured in the η_l term.

Another potential problem is that teacher value-added may be correlated over time for an individual student. If this correlation is positive (i.e. parents request effective teachers every period), the measure of persistence will be biased upwards. One can test the importance of this problem, however, by seeing how the coefficient estimates change when we control for current teacher effectiveness.

3. Data

A. Sample Information

To measure the persistence of teacher-induced learning gains, we use data from the 1998-9 to 2004-5 school years for a mid-size school district located in the western United States.[7] The elemental unit of observation is the individual student, for whom common demographic information such as race, ethnicity, free lunch and special education status, as well as standardized achievement test scores is available. We can

[7] The district has requested to remain anonymous.

track these students over time and link them to each of their teachers, creating a panel of student level observations. This allows us to calculate a value added measure of teacher effectiveness specific to each student to use in our regressions.

In this district, students in grades 1-6 take a set of "Core" exams in reading and math. These multiple-choice, criterion-referenced exams cover topics that are closely linked to the district learning objectives. While student achievement results have not been directly linked to rewards or sanctions until recently, the results of the Core exams are distributed to parents and published annually. Our methodology requires a lagged year of test scores to capture the student's prior performance and a further lag to serve as a potential instrument for long run student ability. This leads us to restrict the sample to grades 3-6 which have twice lagged achievement test scores available.

Because this district uses tracking by ability groups for some mathematics instruction, we restrict math scores to untracked classrooms. Furthermore, sixth grade math classes use different evaluation measures, and are thus excluded from the analysis. Although we use a normalized test score measure, scaled to report standard deviation units relative to the district, as the outcome variable, robustness checks with percentile ranked scores yield similar results.

The summary statistics of Table 1 show that although the Grade 3-6 students in the district are predominantly white (76 percent), there is a reasonable degree of heterogeneity in other dimensions. For example, close to half of all students in the district (44 percent) receive free or reduced price lunch, and about 10 percent have limited English proficiency.[8] Although we do not use teacher characteristics in the

[8] Achievement levels in the district are almost exactly at the average of the nation, with students scoring at the 49th percentile on the Stanford Achievement Test.

analysis, along observable dimensions the teachers constitute a fairly close representation of elementary school teachers nationwide.

B. Estimating teacher value added.

To measure the persistence of teacher-induced learning gains we must first estimate teacher value added. Consider a learning equation of the following form.

(13) $test_{ijt} = \beta test_{it-1} + X_{it}\Gamma + \theta_j + \eta_{jt} + \varepsilon_{ijt}$,

where $test_{it}$ is a test score for individual *i* in period *t*, X_{it} is a set of potentially time varying covariates, θ_j captures teacher value-added, η_{jt} reflects period specific classroom factors that affect performance (e.g. test administered on a hot day or unusually good chemistry between the teacher and students), and ε_{it} is a mean zero residual.

There are two concerns regarding our estimates of teacher value added. The first, discussed earlier, is that the value added measures may be inconsistent due to the non-random assignment of students to teachers. The second is that the imprecision of our estimates may affect the implementation of our strategy. Standard fixed effects estimation of teacher value added rely on test score variation due to classroom specific learning shocks, η_{jt}, as well as student specific residuals, ε_{ijt}. Because of this, the estimation error in teacher value added will be correlated to contemporaneous student achievement and fail to satisfy the necessary exclusion restrictions for consistent instrumental variables identification.

To avoid this problem we estimate the value added of a student's teacher while ignoring the contribution of that student's cohort. In practice we accomplish this by estimating a separate regression for each cell of year by grade student level observations, and recording the teacher value added estimates. In each regression we control for student age, race, gender, free-lunch eligibility, special education placement, limited English proficiency status, class size and school fixed effects. Then for each student we compute an average of his teacher's value added measures across all years in which that student was not in the teacher's classroom.[9] The estimation error of the resulting value added measures will be uncorrelated to unobserved determinants of the reference student's achievement.

Table 2 presents summary measures of these value added metrics. Although they are approximately mean zero by design, the dispersion for our normalized scores is close to that found in previous studies such as Rockoff (2004) and Aaronson, Barrow and Sander (2007). As discussed later, the results of our estimation are robust to various specifications of the initial value added equation.[10] However, as previously suggested, it is likely that non-random sorting of students to teachers will bias our estimates upwards, leading us to overstate persistence.

4. Results

This section presents the results of our estimation of the persistence of teacher value added induced learning. Table 3 considers the baseline case where we examine persistence after one year in a specification with the full student and classroom level controls including race, gender, free lunch eligibility, special education status and limited

[9] This has the added benefit of greatly improving the reliability of the teacher value added measures over simple one year measures (McCafferey, Lockwood and Sass 2008).

[10] More details on robustness checks of teacher value added measures using this dataset can be found in Jacob and Lefgren (2005a) and (2005b)

English status as well as school and year fixed effects. We also control for grade fixed effects and allow the slopes of all covariates and instruments to vary by grade (the coefficient on lagged achievement is constrained to be the same for all students). Instrumental Variables estimates of long run learning persistence use twice lagged test scores and an indicator for a missing twice lagged score as excluded instruments. Estimates of the persistence of teacher value added use the previously calculated value added measures interacted with grade dummies as excluded instruments.

Our estimate of the persistence of knowledge from all sources, $\hat{\beta}_{OLS}$, is 0.66 for reading and 0.62 for math, suggesting that two-thirds of a general gain in student level test scores is likely to persist after a year.[11] In contrast, the estimate of $\hat{\beta}_{LR}$, suggests that variation in test scores caused by prior (long-run) learning is almost completely maintained.

When compared against these baselines, the achievement gains due to a high value added teacher are more ephemeral, with point estimates suggesting that only about one-fifth of the initial gain is preserved after the first year. However, our results also statistically reject the hypothesis of zero persistence at conventional significance levels.[12] For the latter two coefficient estimates, the table also reports the F statistic of the instruments used in the first stage. In all cases the instruments have sufficient power to make a weak instruments problem unlikely.

Table 4 considers the persistence of achievement after two years. The estimation strategy is analogous to that of Table 3, except that the coefficient of interest is now that

[11] This estimate of persistence from all sources is comparable to that of other recent studies such as Todd and Wolpin (2006) and Sass (2006).

[12] Reported standard errors are corrected for classroom level clustering.

of the second lag of student test scores. All instruments are also lagged an additional year. In all cases, most of the gains that persist in the first year continue in the second.[13] In reading, persistence in test score increases from all sources and persistence of gains due to teacher value added drop 6-9 percentage points from their one year levels. Math scores drop by 3-6 percentage points. In both cases the drop in persistence of gains from teacher value added appears to be slightly larger, although not distinguishable statistically.

Long term learning continues to demonstrate nearly perfect persistence. It is slightly surprising that after losing four-fifths of the gains from teacher value added in the first year, students in the next year only lose a few percentage points. This suggests that our data generating model is a good approximation to the actual learning environment in that much of the achievement gain maintained beyond the first year may be permanent. However, most of the overall gain attributed to value added is still a temporary one period increase.

These results are largely consistent with the current evidence on persistence. McCaffery et al (2004) find a persistence of around one-fifth (although quite imprecisely estimated). In their observational data Kane and Staiger (2008) find one year persistence estimates of 0.22-0.25 and two year persistence estimates between 0.12-0.14, although their estimates are more sensitive than ours to the addition of contemporary classroom fixed effects. Furthermore, estimation within their experimental period suggests the possibility of higher rates of persistence for language scores.

[13] There is a sample disparity between the 1 year and 2 year persistence estimates since the latter do not contain third graders due to the need for an additional lag of test scores.

Table 5 presents a series of robustness checks for our estimation of $\hat{\beta}_{VA}$. The primary obstacle to identifying a true measure of the persistence of teacher value added is the possible non-random assignment of students to teachers, both contemporaneously, and in prior years. Although we attempt to deal with this possibility with a value added model and the inclusion of student and peer characteristics in the regression, it is still possible that we fail to account for systematic variation in the assignment of students to teachers. Row (2) of Table 5 presents estimates of the persistence of value added when all controls except for school, grade and year fixed effects are dropped from the regression model. In all cases the coefficient estimates increase, suggesting that there is positive selection on observables. This matches with our priors that the assignment system may favor highly invested parents by assigning their students to better teachers. However, if there exists a positive selection on unobservables that is not controlled for by our estimation strategy, then $\hat{\beta}_{VA}$ is actually an upper bound for the true effect. Thus the most likely identification failure suggests an even lower persistence than we find in Tables 3 and 4.

The remainder of the table suggests that our estimates are quite robust to changes in the regression model. Row 3 adds contemporary classroom fixed effects with only a slight attenuation of the estimated coefficients, suggesting that principals are not likely compensating students for past teacher assignments. The next three rows consider the impact of modifying the procedure for estimating teacher value added measures. The first uses a gains specification as opposed to lag specification of value added while the second further normalizes those gains by the initial score and the third uses only students in the middle of the achievement distribution to calculate teacher value added to minimize the possible influence of outliers. This last check produces a large increase in the coefficient for the two year persistence of math scores. Otherwise all the estimates represent only small deviations from the baseline. The final specification check measures all test performance in percentiles of the district distribution and finds the same substantial persistence pattern.

There seems to be a clear pattern of evidence for small, non-zero levels of teacher value added persistence. However, these measured effects are averages across a heterogeneous population of students. Table 6 considers the degree to which the persistence estimates differ across years, grades and some student characteristics. For each characteristic group we present a chi-squared statistics for a test of the null hypothesis that the coefficients are equal across all groups and the p-value for that test. The first panel considers differences across test years. While the hypothesis of coefficient inequality is formally rejected for the one year math persistence only, there appears to be a cross year pattern for all other test score categories. In general the 1999, 2002-3 and 2005 have measured effects near the baseline, 2004 has effects well above the baseline and 2000 and 2001 have widely ranging estimated effects including some negative estimates. While it is certainly possible that this is due to actual changes in the persistence across years it also seems possible that some of the difference may be due to differences in the test instrument across years.

The second panel considers cross grade differences. In reading we reject the hypothesis that persistence is the same across grades, while we fail to reject this hypothesis for math persistence. The pattern of coefficients is consistent with the case in which the carryover in curriculum from one grade to the next may vary across grades. No matter how good the teacher is, if they are not teaching knowledge that will play a direct role in the next year's exams we will see little persistence. Furthermore the large significant coefficient on the one year reading persistence for fourth graders, and the two year persistence coefficient for fifth graders suggests that the third grade reading curriculum presents greater opportunities for teachers to convey long run knowledge than the curricula of other grades. To the degree that math algorithms tend to have more general long term uses compared to what students may do in a reading class this is not surprising.

The final four panels of the table consider the heterogeneity of persistence across groups of students with different observable characteristics. In all cases, students eligible for free lunch had lower estimated persistence measures than ineligible students, although the difference is only significant for one year math scores. Although this appears to suggest that disadvantaged students derive less persistent benefits from teacher value added, the following two panels suggest the true situation is much more complicated. Minority students, for example, have a statistically significant advantage in measured persistence for reading scores, while limited English proficient students have a negative estimated persistence for teacher value added using math scores, but no such disparity in reading. There is no apparent pattern of differing results across gender groups.

Perhaps the primary claim of the teacher value added literature is that teacher quality matters a great deal for student achievement. This is based on consistent findings of a large dispersion in teachers' ability to influence contemporary student test scores. However, our results indicate that this dispersion and subsequent claims about teacher influence are overstated. Any teacher value added measure over contemporary outcomes conflates variation in short term and long term knowledge. Given that a school's objective is to increase the latter, the variance of teacher ability to meet the goal is substantially less than the teacher value added literature indicates. Note that this does not mean the average level of teacher value added is unimportant, rather that the variation in the distribution of existing teacher value added is less informative than contemporary test gains suggest.

It is also worth noting that our statistical model may capture a number of different rationalizations of knowledge fadeout. The simplest, almost tautological, reason why short run knowledge if different from long run knowledge is short run knowledge is forgotten by students. However, this does not satisfactorily explain why smaller classes might induce a different persistence in knowledge than higher value added teachers. This distinction among programs could be a result of several possible mechanisms. For example, persistence might be a relic of school organization or test structure. School programs with more material in common across grades and exams will look more persistent. Alternatively, low persistence in teacher value added effects could be due to some sort of compensatory teaching, whereby later teachers change their curriculum to address the students who learned less in previous years because they had low value added teachers. Finally, in the presence of accountability programs the low persistence could be an artifact of teachers spending time on subjects with high test specific returns but low long term educational value. In addition to further research comparing persistence across programs, additional inquiry into the mechanisms that drive these differences is needed.

Previous researchers have referenced a counterfactual world where a series of high value added effects for a hypothetical student with a string of good teachers may simply be added together. Given these hypothesized achievement gains, the policy implication for spending programs that attempt to improve teacher value added seems clear. Our results, however, suggest caution in such claims. Clearly, if value added test score gains do not persist over time, adding up consecutive gains is unlikely to measure a meaningful policy alternative. Nevertheless, our results suggest there is some long-run persistence to the gains induced by teacher value added, even if it is small compared to the persistence of test score gains from all sources. It is possible that improving teacher value added will improve the long-run outcomes of students.

5. Conclusion

Our statistical model formalizes the suspicion that all increases in test scores are not equal. Relative to both the variation in test scores generated by all sources and the variation induced by long run learning, test score gains created by a high value added teacher have low persistence. Our estimates suggest that only about one-fifth of the test score gain from a high value added teacher remains after a single year. Given our standard errors we can rule out one year persistence rates above one-third. After two years about one-eighth of the original gain persists. This attrition is observed for both math and reading scores and is robust to several specification checks. Furthermore, the positive selection on observables suggests that our estimates may be overly optimistic.

After decades of pessimism concerning the lack of connection between the measurement of teacher observable characteristics and student achievement, the use of fixed effects value added measures for teachers have led to renewed optimism about the ability to measure, reward and provide incentives for teacher effectiveness. Our results suggest that policies based on short run value added estimates may have disappointing long run results.

The econometric framework we use to measure the persistence of teacher induced learning gains is more broadly applicable. It can be used to the measure the persistence of any educational intervention. Relative to the informal methods previously used, our approach allows simple statistical inference, clear comparison across policies, and clearly relates the empirical results to the assumed data generating process.

References

Aaronson, Daniel, Lisa Barrow, and William Sander. (2007) Teachers and Student Achievement in the Chicago Public High Schools. *Journal of Labor Economics 25* (1), 95-135.

*Andrabi, Tahir, Tishnu Das, Asim I. Khwaja, and Tristan Zajonc (2007) "Do Value-Added Estimates Add Value? Accounting for Learning Dynamics" mimeo.

Barnett, W. S. (1985). Benefit-Cost Analysis of the Perry Preschool Program and Its Policy Implications. *Educational Evaluation and Policy Analysis (7)*. 333-342.

Boardman, A. & Murnane, R. (1979), .Using panel data to improve estimates of the determinants of educational achievement., *Sociology of Education 52* (2), 113-121.

Coleman, James S., et al. (1966). *Equality of educational opportunity*. Washington, DC: U.S. Government Printing Office.

Currie, Janet., and Duncan Thomas. (1995) Does Head Start Make A Difference?, *The American Economic Review* 85 (3), 341-364.

Doran, H. & Izumi, L. (2004). Putting Education to the Test: A Value-Added Model for California., San Francisco: Pacific Research Institute.

Goldhaber, Dan D., and Dominic J. Brewer. (1997). Why don't school and teachers seem to matter? *Journal of Human Resources* 32, no. 3:505–23.

Hanushek, Eric A. (1997). Assessing the effects of school resources on student performance: An update. *Education Evaluation and Policy Analysis* 19:141–64.

Hanushek, Eric A. (2003), .The failure of input-based schooling policies., *Economic Journal* 113, 64-98.

Imbens, Guido W & Angrist, Joshua D, (1994). "Identification and Estimation of Local Average Treatment Effects," Econometrica, Econometric Society, vol. 62(2), pages 467-75, March.

*Jacob, Brian A. and Lars Lefgren. (2005a) “Principals as Agents: Subjective Performance Measurement in Education,” National Bureau of Economic Research Working Paper No. 11463, June 2005.

*Jacob, Brian A. and Lars Lefgren. (2005b). “What Do Parents Value in Education? An Empirical Investigation of Parents’ Revealed Preferences for Teachers.” National Bureau of Economic Research Working Paper No. 11494.

Jacob, Brian A., and Lars Lefgren. (2004). Remedial Education and Student Achievement: A Regression-Discontinuity Analysis. *Review of Economics and Statistics (86)*. 226-44.

Jacob, Brian A. and Steven Levitt, (2003). Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating. *Quarterly Journal of Economics 118*. 843-77.

*Kane, Thomas J., and Douglas O. Staiger. (2008). Are Teacher-Level Value-Added Estimates Biased? An Experimental Validation of Non-Experimental Estimates. Mimeo March 17.

Kane, Thomas J., and Douglas O. Staiger. (2002). The promises and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives* 16, no. 4:91–114.

Krueger, Alan B., and Diane 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. *Economic Journal (111).* 1-28.

*McCaffrey, Daniel F., J. R. Lockwood, and Tim R. Sass. (2008) The Inter-temporal Stability of Teacher Effect Estimates. Mimeo.

McCaffrey, Daniel F., J. R. Lockwood, Daniel Koretz, Thomas A. Louis, and Laura Hamilton (2004) "Models for Value-Added Modeling of Teacher Effects" *Journal of Educational and Behavioral Statistics*, Vol. 29, No. 1, Value-Added Assessment Special Issue., Spring, pp. 67-101.

Nye, Barbara, Larry V. Hedges, and Spyros Konstantopoulos. (1999). The Long-Term Effects of Small Classes: A Five-Year Follow-Up of the Tennessee Class Size Experiment. *Educational Evaluation and Policy Analysis (21).* 127-142

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. (2005). Teachers, schools, and academic achievement. *Econometrica* 73, no. 2:417–58.

Rockoff, Jonah E., (2004) "The impact of individual teachers on student achievement: evidence from panel data," *American Economic Review*. 247-252.

*Rothstein, Jesse. (2007). Do Value Added Models add value? Tracking, Fixed Effects and Causal Inference. Mimeo. November 20.

Sass, T. (2006). Charter schools and student achievement in Florida. *Education Finance and Policy* 1(1), 91-122.

Todd, P. & Wolpin, K. (2003), .On the Speci.cation and Estimation of the Production Function for Cognitive Achievement., *Economic Journal 113*, 3-33.

Todd, P. &Wolpin, K. (2006), .The Production of Cognitive Achievement in Children: Home, School and Racial Test Score Gaps., Philadelphia, PA: University of Pennsylvania, *PIER Working Paper* pp. 4-19.

Table 1: Summary Statistics

Variable	Mean (Std. dev.)
Normalized Reading Score	-0.018 (0.985)
Normalized Math Score	0.026 (0.976)
Reading Percentile Score	0.476 (0.280)
Math Percentile Score	0.489 (0.280)
Student Fraction Male	0.505 (0.500)
Student Fraction Free Lunch	0.436 (0.496)
Student Fraction Minority	0.239 (0.427)
Student Fraction Special Ed.	0.083 (0.276)
Student Fraction Limited English	0.101 (0.301)
Student Age	10.921 (1.157)
Grade 4	0.263 (0.440)
Grade 5	0.246 (0.430)
Grade 6	0.217 (0.412)

Notes: Test scores are normalized relative to the standard deviation for all students in the district.

Table 2: Summary of Teacher Value Added Measures

Measure	Mean (Std. Dev.)
Reading normalized value added of student's Teacher (t-1)	0.016 (0.212)
Math normalized value added of student's Teacher (t-1)	0.027 (0.294)
Reading normalized value added of student's Teacher (t-2)	0.011 (0.229)
Math normalized value added of student's Teacher (t-2)	0.009 (0.304)

Notes: Test scores are normalized relative to the standard deviation for all students in the district.

Table 3: Measuring the One Year Persistence of Achievement

	Reading			Math		
	$\hat{\beta}_{OLS}$	$\hat{\beta}_{LR}$	$\hat{\beta}_{VA}$	$\hat{\beta}_{OLS}$	$\hat{\beta}_{LR}$	$\hat{\beta}_{VA}$
Prior Year Achievement	0.66**	0.98**	0.22**	0.62**	0.98**	0.19**
Coefficient	(0.009)	(0.02)	(0.06)	(0.01)	(0.02)	(0.06)
F-Statistic of Instruments	--	1,412	48	--	839	65
[p-value]		[0.00]	[0.00]		[0.00]	[0.00]
Observations	18,240	18,240	18,240	14,182	14,182	14,182
R-Squared	0.59	0.51	0.44	0.51	0.41	0.36

Notes: Reported standard errors in parentheses correct for clustering at the classroom level. ** indicates 5% significance, * 10% significance.

Table 4: Measuring the Two Year Persistence of Achievement

	Reading			Math		
	$\hat{\beta}_{OLS}$	$\hat{\beta}_{LR}$	$\hat{\beta}_{VA}$	$\hat{\beta}_{OLS}$	$\hat{\beta}_{LR}$	$\hat{\beta}_{VA}$
Two Year Prior	0.60**	0.95**	0.13**	0.59**	0.97**	0.13
Achievement Coefficient	(0.01)	(0.03)	(0.06)	(0.02)	(0.04)	(0.08)
F-Statistic of Instruments	--	961	55	--	439	63
[p-value]		[0.00]	[0.00]		[0.00]	[0.00]
Observations	10,216	10,216	10,216	7,104	7,104	7,104
R-Squared	0.54	0.44	0.36	0.49	0.37	0.31

Notes: Reported standard errors in parentheses correct for clustering at the classroom level. ** indicates 5% significance, * 10% significance.

Table 5: Robustness Checks

		Reading		Math	
		1 Year Persistence	2 Year Persistence	1 Year Persistence	2 Year Persistence
(1)	Baseline	0.22** (0.06)	0.13** (0.06)	0.19** (0.06)	0.13 (0.08)
(2)	Controlling *Only* for Grade, School, and Year in Second Stage	0.32** (0.06)	0.23** (0.06)	0.22** (0.06)	0.19** (0.08)
(3)	Controlling for Classroom Fixed Effects in Second Stage	0.19** (0.05)	0.14** (0.06)	0.12** (0.05)	0.11* (0.07)
(4)	Value-Added Estimated Using Achievement Gains	0.15** (0.07)	0.11 (0.07)	0.08 (0.07)	0.08 (0.09)
(5)	Value-Added Estimated Using Achievement Gains Normalized by Initial Score	0.16** (0.06)	0.10 (0.07)	0.16** (0.07)	0.14* (0.08)
(6)	Value-Added Estimated Using Students in Middle of Achievement Distribution	0.15** (0.07)	0.14** (0.06)	0.18** (0.07)	0.24** (0.08)
(7)	Test Performance Measured in Percentiles of District Performance	0.17** (0.07)	0.14** (0.05)	0.18** (0.05)	0.14** (0.06)

Notes: Reported standard errors in parentheses correct for clustering at the classroom level. ** indicates 5% significance, * 10% significance.

Table 6: Heterogeneity of Persistence of Teacher Induced Achievement

	Reading		**Math**	
	1 Year Persistence	2 Year Persistence	1 Year Persistence	2 Year Persistence
Baseline	0.22**	0.13**	0.19**	0.13
	(0.06)	(0.06)	(0.06)	(0.08)
		Year		
1999	0.33**	--	0.39**	--
	(0.10)		(0.08)	
2000	-0.08	0.30**	-0.21	0.25**
	(0.18)	(0.14)	(0.19)	(0.10)
2001	0.13	0.00	0.08	0.00
	(0.10)	(0.17)	(0.13)	(0.14)
2002	0.31**	0.15	0.01	0.16
	(0.14)	(0.10)	(0.13)	(0.15)
2003	0.30**	0.07	0.32**	0.13
	(0.12)	(0.19)	(0.09)	(0.15)
2004	0.47**	0.16	0.32**	0.27**
	(0.12)	(0.13)	(0.14)	(0.13)
2005	0.37**	0.16	0.41**	0.02
	(0.15)	(0.17)	(0.13)	(0.22)
χ^2 Equal Coefficients	9.22	2.07	15.23	3.46
[P-value]	[0.16]	[0.84]	[0.02]	[0.63]
		Grade		
Third	0.14	--	0.18	--
	(0.12)		(0.13)	
Fourth	0.38**	0.08	0.23**	0.14
	(0.09)	(0.16)	(0.08)	(0.13)
Fifth	-0.19	0.28**	0.06	0.12
	(0.16)	(0.07)	(0.18)	(0.09)
Sixth	0.41**	-0.06	--	--
	(0.15)	(0.12)		
χ^2 Equal Coefficients	11.76	6.89	0.76	0.02
[P-value]	[0.01]	[0.03]	[0.69]	[0.88]
		Free Lunch Status		
No	0.28**	0.21**	0.33**	0.23**
	(0.07)	(0.09)	(0.07)	(0.09)
Yes	0.19**	0.07	0.03	0.00
	(0.09)	(0.10)	(0.09)	(0.12)
χ^2 Equal Coefficients	0.53	1.15	6.16	2.40
[P-value]	[0.47]	[0.28]	[0.01]	[0.12]

		Minority Status		
No	0.13	0.07	0.20**	0.15
	(0.08)	(0.08)	(0.08)	(0.09)
Yes	0.45**	0.28**	0.19*	0.04
	(0.08)	(0.11)	(0.10)	(0.12)
χ^2 Equal Coefficients	8.41	2.45	0.01	0.63
[P-value]	[0.00]	[0.12]	[0.94]	[0.43]
		Limited English Proficiency		
No	0.21**	0.14**	0.22**	0.18**
	(0.07)	(0.07)	(0.07)	(0.08)
Yes	0.27**	0.06	-0.02	-0.27
	(0.12)	(0.20)	(0.17)	(0.20)
χ^2 Equal Coefficients	0.19	0.17	1.63	3.74
[P-value]	[0.67]	[0.68]	[0.20]	[0.05]
		Gender		
Female	0.22**	0.11	0.19**	0.19*
	(0.08)	(0.09)	(0.08)	(0.10)
Male	0.23**	0.16*	0.20**	0.07
	(0.08)	(0.08)	(0.08)	(0.10)
χ^2 Equal Coefficients	0.03	0.12	0.03	0.63
[P-value]	[0.85]	[0.73]	[0.86]	[0.43]

Notes: Reported standard errors in parentheses correct for clustering at the classroom level. ** indicates 5% significance, * 10% significance. Figures in brackets are p-values for the chi-square test of coefficient inequality across groups.