The Persistence of Teacher-Induced Learning[*]

Brian A. Jacob

Lars Lefgren

David P. Sims

July 2009

Abstract

This paper constructs a statistical model of learning that suggests a systematic way of measuring the persistence of treatment effects in education. This method is straightforward to implement, allows for comparisons across educational treatments, and can be related to intuitive benchmarks. We demonstrate the methodology using student–teacher linked administrative data for North Carolina to examine the persistence of teacher quality. We find that teacher-induced learning has low persistence, with three-quarters or more fading out within one year. Other measures of teacher quality produce similar or lower persistence estimates.

I. 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 suggests that the positive effects of educational interventions may fadeout over time (see the work on Head Start by Currie and Thomas (1995), and others). Failure to accurately account for this fadeout can dramatically change the assessment of a program's impact and/or its cost effectiveness.

Unfortunately, work on measuring the persistence of educational program impacts has not received much emphasis in the applied microeconomics literature, particularly in the area of teacher effectiveness which has been a focus of comparatively detailed attention among researchers and policymakers. Indeed, a number of districts and states are experimenting with ways to advance the use of teacher characteristics, including statistically derived "value-added" measures in the design of hiring, certification, compensation, tenure, and accountability policies. An oft-cited claim is that matching a student with a stream of high value-added teachers (one standard deviation above the average teacher) for five years in a row would be enough to completely eliminate the achievement gap between poor and non-poor students (Rivkin, Hanushek, and Kain 2005). This prognosis fails, however, to mention the underlying assumption of perfectly persistent teacher effects, that is effects with identical long- and short-run magnitudes.

This paper advances the persistence literature by introducing a framework for estimating

and comparing the persistence of treatment effects in education across policy options. To begin, we present a simple model of student learning that incorporates permanent as well as transitory learning gains. 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 local average treatment effect (Imbens and Angrist 1994). We illustrate our method by estimating the persistence of teacher-induced learning gains with multiple measures of teacher quality, though the method generalizes to other educational interventions. Using the student-teacher linkages available in administrative data from North Carolina, we construct measures of teacher effectiveness, including observable teacher correlates of student achievement such as experience and credentials (Clotfelter, Ladd, and Vigdor 2007) as well as statistically derived measures of teacher value-added.

Our resulting estimates suggest that teacher-induced variation in both math and reading achievement quickly erodes. Our point estimates indicate a one-year persistence of teacher value-added effects near one-fourth for math and one-fifth for reading. Furthermore, we find that the estimated persistence of non-test based measures of teacher effectiveness is, at best, equal to that of value-added measures. These results are robust to a number of specification checks and suggest that depreciation of a similar magnitude applies to different student racial, gender and socioeconomic groups. Further estimates suggest that only about one-sixth of the original student gains from a high value-added teacher persist over two years. We further discuss what these estimates can tell us about the relative importance of three fadeout mechanisms: forgetting, compensatory investment, and future learning rates.

In general, our evidence suggests that even consistent estimates of single-period teacher quality effects drastically overstate the relevant long-run increase in student knowledge. Our

results highlight the potential importance of incorporating accurate persistence measures in educational policy evaluation and suggest a comparative framework for implementation.

This paper focuses on the persistence of teacher effects, an issue which is distinct from the potential bias in teacher value-added estimates due to omitted variables or non-random assignment of teachers. However, we are still concerned about the potential effects of this bias on our estimates, and we discuss this issue in detail below. We believe that at a minimum our estimates still present a useful *upper* bound to the true persistence of teacher effects on student achievement.

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

II. A Statistical Model

This section outlines a 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 consider other educational interventions.

A. Base Model

In order to control for past student experiences, education researchers often employ empirical strategies that regress (mean zero measures of) current achievement on lagged

achievement, namely

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

with the common result that the OLS estimate of β is less than one. This result is typically

given one of two interpretations: Either the lagged achievement score is measured with error due

to factors such as guessing, test conditions, or variation in the set of tested concepts, or the

coefficient represents the constant depreciation of knowledge over time.

In order to explore the persistence of knowledge, it is useful to specify the learning

process underlying these test scores. To begin, suppose that true knowledge in any period is a

linear combination of what we describe as "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.[1] 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 decays entirely by the next period.[2] 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 δ, the persistence of long-term 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 likely form part of long-term knowledge while the latter would be quickly forgotten. Thus it is likely that a given teacher affects both long and short-term knowledge, though perhaps to different degrees.

It is worth noting that variation in knowledge due to measurement error is observationally equivalent to variation due to the presence of short-run (perfectly depreciable) knowledge in this model, even though these may reflect different underlying mechanisms. For example, both a teacher cheating on behalf of students and 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. Similarly, a student always forgetting material of a particular nature would appear as short-term knowledge.[3] Consequently, our persistence estimates do not directly distinguish between short-run knowledge that is a consequence of limitations in the ability to measure achievement and short-run knowledge that would have real social value if the student retained it.

In most empirical contexts, the researcher only observes the total of long- and short-run knowledge, $Y_t = y_{l,t} + y_{s,t}$, as is the case when one can only observe a single test score. 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.[4] It is then straightforward to show that when considering this composite test score, Y_t, in the typical "value-added" regression model given by equation (1), the OLS estimate of β converges to:

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

Thus, OLS identifies the persistence of long-run knowledge multiplied by the fraction of variance in total knowledge attributable to long-run knowledge. In other words, one might say that 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 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:

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

where lagged achievement is regressed on twice-lagged achievement. 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:

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

which is the persistence of long-run knowledge. Our estimates suggest that this persistence is close to one.

Now consider what happens if we instrument lagged knowledge, Y_{t-1}, with the lagged teacher's contribution (value-added) to total lagged knowledge. The first stage is given by:

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

where the teacher's total contribution to lagged knowledge is a combination of her contribution to long- and short-run lagged knowledge, $\Theta_{t-1} = \theta_{l,t-1} + \theta_{s,t-1}$. In this case, the second stage estimate, which we refer to as $\hat{\beta}_{VA}$ converges to:

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

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 her contribution to short-term knowledge. Given our estimates of δ, the persistence of long-run knowledge, are roughly equal to one, $\hat{\beta}_{VA}$ approximates 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.[5] 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 while $\hat{\beta}_{LR}$ measures the persistence of long-run knowledge, that is achievement differences due to prior knowledge.

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.[6] In this case, instrumenting lagged observed knowledge with lagged treatment assignment yields an estimator which converges to the following:

$$(10) \quad plim\left(\hat{\beta}_{TREAT}\right) = \delta \frac{\gamma_1}{\gamma_1 + \gamma_s}.$$

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.

A standard approach to estimating the persistence of treatment effects is to simply compare the ratio of coefficients from separate treatment effect regressions at different points in the future. For example, one might estimate the impact the impact of a child's current fifth grade teacher on her contemporaneous fifth grade test scores, and then in a second regression estimate the impact of the child's former (in this case fourth grade) teacher on her fifth grade test scores. The ratio of the teacher coefficient from the second regression to the analogous coefficient in the first regression provides a measure of the one-year persistence of the teacher effect.

While this approach does provide a measure of persistence, our approach has a number of advantages over the informal examination of coefficient ratios. First, it is provides a straightforward way to both compute estimates and conduct inference on persistence measures through standard t- and F-tests.[7] Second, the estimates of $\hat{\beta}_{LR}$ and $\hat{\beta}_{OLS}$ serve as intuitive

benchmarks that allow an understanding the relative importance of teacher value-added as opposed to test scaling effects. They allow us to examine the persistence of policy induced learning shocks relative to the respective effects of transformative learning and a "business as usual" index of educational persistence. Finally, the methodology can be applied to compare persistence among policy treatments including those that that may be continuous or on different scales such as hours of tutoring versus number of students in a class.

B. Extensions

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

$$(11) \quad plim\left(\hat{\beta}_{VA}\right) = \delta \frac{\sigma_{\theta_l}^2 + \text{cov}\left(\theta_l, \theta_s\right)}{\sigma_{\theta_l}^2 + \sigma_{\theta_s}^2 + 2\,\text{cov}\left(\theta_l, \theta_s\right)} = \delta \frac{\text{cov}\left(\theta_l, \Theta\right)}{\sigma_{\Theta}^2}.$$

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 the teacher's contribution to total measured knowledge.

Another interesting consequence of relaxing this independence assumption is that β_{VA} need not be positive. In fact, if $\text{cov}\left(\theta_l, \theta_s\right) < -\sigma_{\theta_l}^2$, β_{VA} will be negative. This can only be true if $\sigma_{\theta_l}^2 < \sigma_{\theta_s}^2$. 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.[8] [9]

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 $\mathrm{cov}\left(\theta_{l,t}, \eta_{l,t}\right) \neq 0$, as would occur if school administrators systematically allocate children with unobserved high learning 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:

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

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 and contemporaneous classroom fixed effects in the second-stage regression, which would be captured in the η_l term. Indeed, the inclusion of student covariates reduces the estimated persistence measure, which suggests that such a simple positive selection story is the most likely and our persistence measures are overestimates of true persistence. On the other hand, Rothstein (2009) argues that the matching of teachers to students is based in part on transitory gains on the part of students in the previous year. This story might suggest that we

underestimate persistence, as the students with the largest learning gains in the prior year are assigned the least effective teachers in the current year.

Another potential problem is that teacher value-added may be correlated over time for an individual student. If this correlation is positive, perhaps because motivated parents request effective teachers every period, the measure of persistence will be biased upwards. We explore this prediction by testing how the coefficient estimates change when we omit our controls for student level characteristics. In our case the inclusion of successive levels of control variables monotonically reduces our persistence estimate, as the positive sorting story would predict.

III. Background

A. Teacher Quality and Value-Added Measures

There are a number of recent studies suggesting an important role for teacher quality in elementary and secondary education based on its effects on contemporaneous test scores. Their prevalence suggests that this is an important area in which to examine the role of long-run versus short-run learning as described in our model. One branch of these studies indicate that some observable teacher characteristics such as certification, experience, and principal evaluations may have small but statistically significant effects on student test scores (Kane, Rockoff, and Staiger 2008; Clotfelter, Ladd, and Vidgor 2006; Jacob and Lefgren 2007). Another branch uses sophisticated empirical models to attempt to isolate an individual teacher's whole contribution to student test scores. These latter 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 around 0.1 standard deviations. Aaronson, Barrow, and Sander (2007) find similar results using high school data. In comparison, it would require a 4-5 student decrease in class size to

achieve the same effect as a one standard deviation increase in teacher value-added (Angrist and Lavy 1999).

This research has inspired proposals that seek to use value-added metrics to evaluate the effectiveness of classroom teachers for compensation or tenure purposes (Doran and Izumi 2004, McCaffrey et al. 2004). Given the poor record of single year test scores (Kane and Staiger 2002) or even principal evaluations (Jacob and Lefgren 2008) 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.

At the same time, a number of recent studies (Andrabi et al. 2008; McCaffrey et al. 2004; Rothstein 2009; Todd and Wolpin 2003, 2006) highlight the strong assumptions of the most commonly used value-added models, suggesting that 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 (2009) 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 against these criticisms than gains models.

B. Prior Literature

As Todd and Wolpin (2003) note, much of the early research on teacher value-added fails to explicitly consider the implications of imperfect persistence. For example, the Rivkin, Hanushek, and Kain (2005) scenario of "five good teachers" assumes perfect persistence of student gains due to teacher quality. As a result, these studies imply that variation in test score increases due to policy changes will have long-run consequences equivalent to those of test score increases that come from increased parental investment or innate student ability.

The first paper to explicitly consider the issue of persistence in the effect of teachers on student achievement was a study by McCaffery et al. (2004). Although their primary objective is to test the stability of teacher value-added models to various modeling assumptions, they also provide parameter estimates from a general model that explicitly considers the one- and two-year persistence of teacher effects on math scores for a sample of 678 third through fifth graders from five schools in a large school district. Their results suggest one year persistence of 0.2 to 0.3 and two-year persistence of 0.1. However, due to the small sample the standard errors on each of these parameter estimates was approximately 0.2.

In a later article, Lockwood et al. (2007) produce a Bayesian formulation of this same model which they use to estimate persistence measures for a cohort of approximately 10,000 students from a large urban school district over five years. Using this computationally demanding methodology, they produce persistence estimates that are in all cases below 0.25 with relatively small confidence intervals that exclude zero and appear very similar for both reading and mathematics. They also note that use of models which assume perfect persistence produce significantly different teacher value-added estimates.

More recently, a contemporary group of teacher value-added studies have emerged that

recognize the importance of persistence. For example, Kane and Staiger (2008) use a combination of experimental and non-experimental data from Los Angeles to examine the degree of bias present in value-added estimates due to non-random assignment of students to teachers. They note that coefficient ratios taken from their results imply a one-year math persistence of one-half and a language arts persistence of 60-70 percent. When they expand their sample to include a more representative group of students and control for additional student characteristics, their persistence estimates drop to near one-fourth. Similarly, Rothstein (2009) mentions the importance of measuring fadeout and presents evidence of persistence effects for a particular teacher around forty percent. Carrell and West (2008) present evidence that more experienced university professors at the Air Force academy induce lower but more persistent variation in student learning.

In summary, while the recent teacher value-added literature has come to recognize the need to account for persistence, it and the broader education production literature still lack a straightforward, systematic way to test hypotheses about persistence and to make cross-program persistence comparisons. Persistence is usually inferred as the informal ratio of coefficients from separate regressions, abstracting from the construction and scaling of the particular exam scores. This seems to be an important omission given previous research suggesting decay rates for educational interventions that vary widely across programs; from long term successes such as the Tennessee class size experiment (Nye, Hedges, and Konstantopoulos 1999; Krueger and Whitmore 2001) or the Perry preschool project (Barnett 1985), to programs with no persistent academic effects such as Head Start (Currie and Thomas 1995) or grade retention for sixth graders (Jacob and Lefgren 2004).

C. Interpreting Persistence

Our measure of persistence reflects three different mechanisms. First, students may forget information or lose skills that they acquired as a result of a particular teacher or intervention. Second, students or schools may engage in potentially endogenous subsequent investments, which either mitigate or exacerbate the consequences of assignment to a particular teacher or intervention. Third, our persistence measure depends on how the knowledge learned from a particular teacher influences student learning of new material.

The use of different performance measures across years will also influence the interpretation of the persistence measure, though we view this not as a separate mechanism, but rather as a factor that affects the weights placed on the three mechanisms described above. Consider, for example, differences in test content across years. To the extent that the knowledge and skills involved in Geometry and Algebra are largely distinct, then the effect of an excellent Algebra teacher may appear to fadeout in the following year when the student is tested only in geometry. In this case, observed persistence will primarily reflect the usefulness of the past intervention in learning new knowledge. In an extreme case, it is possible that each grade's test represents largely unique subject matter, so that persistence of test scores could appear low while an excellent teacher in each grade could raise contemporary learning valued by the labor market. However, given that our analysis focuses on elementary school math and reading, in which there is considerable overlap from year to year, this extreme case seems unlikely. Hence, in our analysis, we believe that persistence will largely reflect the first two mechanisms, forgetfulness and compensatory investments.

Changes over time in the statistical properties of a test, such as variance, may also affect the interpretation of a persistence estimate. If observed performance captures neither a cardinal measure of performance nor a uniformly measured rank, it is impossible to arrive at a unique,

interpretable estimate of persistence. For example, suppose the test metric functions as an achievement ranking but the variance of the performance metric differs across years. In such cases, the observed persistence measure will reflect both the fadeout of teacher-induced changes in rank as well as the cross-year heteroskedasticity of achievement measures. Whatever the true persistence in knowledge, the test scale can be compressed or stretched to produce any desired estimate of fadeout.

This might lead some observers to discount the usefulness of persistence measures in general or reject those estimates they find disagreeable. To us, however, this potential sensitivity of observed persistence to test scale effects underscores the importance of establishing baseline measures of the general persistence of knowledge to which the persistence of teacher-induced knowledge can be compared. In this study, we present two such benchmarks, which allow us to compare the persistence of teacher-induced learning to more familiar sources of variation.

One final concern with interpreting our persistence measures involves the possibility of test manipulation. For example, Jacob and Levitt (2003) document a number of cases of teacher cheating, which led to large observed performance increases in one year which did not persist in subsequent years. Similarly, Carrell, and West's (2008) finding that teacher effects from one year of class can be negatively correlated with future years' exam scores is likely explained by contrasting just such a short-run strategic focus on the part of some teachers with a forward-looking approach on the part of other instructors. Such behaviors would manifest themselves in low observed persistence that one might ascribe to poor test measurement – that is one might argue that there was no true learning in such cases in the first place.

IV. Data

A. The Sample

To measure the persistence of teacher-induced learning gains, we use a data set derived from North Carolina school administrative records maintained by the North Carolina Education Research Data Center. The primary data consists of student-year observations for all students in third through sixth grades in the state from 1997-2004.[10] During this time period North Carolina required end of course standardized exams for all these students in both reading and mathematics that were closely aligned to the state's learning standards.

We follow the practice of earlier researchers (Clotfelter, Ladd, and Vigdor 2007) by standardizing the North Carolina scores to reflect standard-deviation units relative to the state average for that grade and year. Thus, our resulting persistence measure captures the degree to which a teacher's ability to change a student's rank (measured in standard deviations) in the achievement distribution is manifest in the student's rank in subsequent years. This relative measure not only allows for comparability of our results with the prior literature, but also captures the effect of a policy on the ranking of a student at some future date such as entry into the labor market. We also show in our robustness checks that our results are robust to the use of scaled scores, which are designed to approximate an absolute measure of learning.

These student-year records can be matched to personnel record data for classroom teachers that contain information on teacher experience, credentials, and certification. We follow the algorithm described in detail in Clotfelter, Ladd, and Vigdor (2007) to match teachers to students.[11] This allows an approximately 79 percent match success rate for a student's prior year teacher (for whom we wish to calculate persistence). Because the most accurate matching of students to teachers is only possible for third through fifth grade students in this data, and because we require one year of lagged test scores as an instrument to capture long-run learning persistence, we calculate value-added over the set of fourth and fifth grade teachers, and measure

outcomes for these students in fifth and sixth grades. [12] Beyond the matching error rate, the measurement of teacher experience may be a concern as it is calculated as the years for which the teacher is given credit on the salary schedule, (whether those years were in North Carolina or not) a potentially noisy measure of true experience.

Table 1 reports summary statistics including the basic demographic controls for student race, ethnicity, free lunch and special education status available in the data. While the North Carolina sample is close to the national average in free lunch eligibility (44 percent compared to 42 percent nationally), it actually has smaller than average minority enrollments, comprised mainly of African American students, and has only a small percentage of non-native English speakers. About one-eighth of students in North Carolina have a novice teacher (a teacher in his or her first two years of teaching), and a relatively large proportion of teachers, almost ten percent, receive national board certification at some point in the sample period. This latter figure is likely driven by the twelve percent salary premium attached to this certification.

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 (for example a test administered on a hot day or unusually good match quality 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 relies 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 that does not incorporate information from that student's cohort. Specifically, for each year by grade set of student observations we estimate a separate regression of the form:

(14) $test_{ijy} = \beta test_{iy-1} + X_{iy}\Gamma + \varphi_{jy} + \upsilon_{ijy}$,

where indexing of student and teacher remains as above and y now indexes year. X is a vector of control variables including the student's age, race, gender, free-lunch eligibility, special education placement, and limited English proficiency status. Then for each student i with teacher j in year t 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, but in the same school.

(15) $\theta_{ijt} = \sum_{y \neq t} \varphi_{jy}$

Consider, for example, a teacher who taught in school A from 1995-1999 and in school B from 2000-2003. For a student in the teacher's class in 1996, we use a value-added measure that incorporates data from that teacher's classes in 1995 and 1997-1999. For a student in the teacher's class in 2000, we use a value-added measure incorporating data from 2001-2003. [13] The estimation error of the resulting value-added measures will be uncorrelated to unobserved classroom-specific determinants η_{jt} of the reference student's achievement. As discussed later,

the results of our estimation are robust to various specifications of the initial value-added

equation.

Our second-stage equation for estimating the persistence of teacher value-added then

becomes:

(16) $test_{ijt+1} = \beta_{VA} test_{it} + X_{it+1}\Gamma + \theta_{jt+1} + \tau_{t+1} + \gamma_g + \varepsilon_{ijt+1}$

where θ_{ijt} from equation (15) serve as the excluded instruments in the first stage. The

specification includes the above mentioned student level controls as well as grade, year, and

contemporary classroom (teacher) fixed effects. In our second stage, we include classroom fixed

effects (which subsume school fixed effects). Thus, our estimation relies exclusively on

variation in teacher quality within a school.

V. Results

This section presents our estimates of the persistence of teacher-induced learning. The

first three columns of Table 2 consider the baseline case of one-year persistence. Our estimate of

$\hat{\beta}_{OLS}$ comes from a regression of contemporary test score on prior test score and student

demographics. Due to the presence of demographic controls, these estimates differ subtly from

the analogous measure detailed in our statistical model. Namely, the estimates shown in Table 2

capture only the persistence due to sources of variation orthogonal to the included demographic

controls. In practice, however, our estimates of 0.68 for reading and 0.71 for math are nearly

identical to estimates from models that exclude demographics. In addition, the general result that

around two-thirds of the variation in student level test scores is likely to persist after a year is a

benchmark figure confirmed by other studies.[14] Instrumental variables estimates of long-run

learning persistence, $\hat{\beta}_{LR}$, use twice lagged test scores and an indicator for a missing twice lagged

score as the excluded instruments. The estimates suggest that variation in test scores caused by

prior (long-run) learning is almost completely persistent after one year, with estimated values a little below one in both cases.[15]

When compared against these baselines, the achievement differences due to a high value-added teacher, $\hat{\beta}_{VA}$, are more ephemeral. The point estimates suggest that between 0.20-0.27 of the initial test score variation is preserved after the first year. While we statistically reject the hypothesis of zero persistence, the effects are significantly lower than either benchmark. For the instrumental variables 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.[16]

While the simple model presented in section 2 assumes a specific decay process for knowledge, we recognize that this is unlikely to be a complete description of the depreciation of learning. Thus, the final three columns of Table 2 expand the analysis to consider the persistence of achievement after two years. The estimation strategy is analogous to the previous specification, except that the coefficient of interest comes from the second lag of student test scores. All instruments are also lagged an additional year. In all cases, the two-year persistence measures are smaller than the one-year persistence measures. In reading, persistence measures in test score variation due to teacher value-added drop only two percentage points from their one-year levels while math scores appear to lose a third of their one-year persistence.

It may seem slightly surprising that after the observed erosion of the majority of value-added variation in a single year, students in the next year lose a much smaller fraction. This suggests that our data-generating model may be a good approximation to the actual learning environment in that much of the achievement gain maintained beyond the first year is permanent. Alternatively, it may be that even conditional upon our covariates, our measure of persistence

still reflects unobserved heterogeneity in the permanent, unobserved ability of the student. In any event, the large majority of the overall gain commonly attributed to value-added is a temporary one-period increase.

These results are largely consistent with the published evidence on persistence presented by McCaffery et al. (2004) and Lockwood et al. (2007). Both find one- and two-year persistence measures between 0.1 and 0.3. However, our estimates are smaller than those of contemporary papers by Rothstein (2009) and Kane and Staiger (2008), which both suggest one-year persistence rates of 0.4 or greater. As mentioned earlier, different persistence estimates within this band may reflect a different weighting of fadeout mechanisms as well as different scaling issues across outcome measures. Thus perhaps the most important contrast in the Table is the persistence of teacher value-added in relation to the benchmarks. If we think of $\hat{\beta}_{LR}$ as transformative learning, we find that teacher value-added is only one-fourth to one-fifth as persistent, as well as less than half as persistent as the average inputs mark represented by $\hat{\beta}_{OLS}$.

Table 3 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 possibility of 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 characteristics in the regression, it is still possible that we fail to account for systematic variation in the assignment of students to teachers. The first three rows of the table show that relaxing our control strategy results in increases to the estimated persistence. For example, omitting classroom effects (row 3) from the regression leads to significant increases in the estimated persistence, while dropping all other student controls leads the coefficients to increase by a third or more (row 2). This demonstration of positive selection on observables is consistent

with our belief that our $\hat{\beta}_{VA}$ estimate reflects an upper bound in the face of likely positive selection on unobservables. Thus the most likely identification failure suggests an even lower persistence than we find in Table 2.

We have estimated our model using a representation of teacher value-added that controls for lagged student achievement as it may avoid many of the problems associated with the value-added models denominated in terms of test score gains. However, the gains formulation remains popular and rows (4) - (5) show the robustness of our estimates to the use of a gains model in calculating our teacher value-added measures. In the case of math, the switch to a gains specification of value-added has no meaningful effect and in the case of reading it serves only to decrease the estimated persistence by 25-50 percent.

To this point, we have only been measuring the average persistence of teacher-induced test score variation without considering whether the effects are symmetric with respect to the sign of the shocks. Given that persistence measures might be driven by students at the bottom catching up due to non-random school interventions, it seems important to examine the symmetry of the persistence effects. In other words, we wish to see whether the test score consequences of having an uncommonly bad teacher are more or less lasting than the benefits of having an exceptionally good teacher. Rows (6) – (7) of Table 3 show the comparison between effects at the top and bottom of the respective distributions. In all cases, we are unable to reject equal persistence values for both sides of the teacher distribution, though the point estimates are larger for the lower tail.[17]

In the final two rows of Table 3, we examine the sensitivity of our estimates to the scaling of the exam. In row (8), we measure test performance as the achieved percentile within state*year. In row (9), we use the scale score measures, which are commonly treated as if they

possessed a cardinal interpretation. Despite ex-ante concerns that the results may differ, they are surprisingly similar to our baseline suggesting that our finding is not sensitive to the choice among a variety of sensible academic performance measures. Furthermore, examination of the $\hat{\beta}_{LR}$ benchmark for scaled scores finds an estimate of 0.96 for both reading and math, quite close to the standardized benchmark.

Since conclusions concerning the persistence of teacher quality might depend on the heterogeneity of persistence across different groups of students, Table 4 shows how persistence estimates differ across measurable student characteristics as well as year, grade, and teacher experience. For reading scores, there is no evidence of statistically significant heterogeneity of persistence effects across grade, gender, race, test year or free lunch status. For math scores, on the other hand, there appears to be statistically significant differences in persistence across all the above groups except gender. However, the difference in actual magnitudes is small, with a range between 0.03-0.05 for all categories except test year. This uniformity across student groups suggests our measure may be capturing a common effect. The final panel of the table suggests that the persistence of teacher effects is not meaningfully different for experienced versus inexperienced teachers.

For the specifications reported in Table 3 and 4 it is also possible to estimate our benchmark $\hat{\beta}_{LR}$. The results are quite consistent across specifications with estimates always in the 0.95-0.99 range for reading scores and the 0.93-0.97 range for math scores. Given this small range we have chosen not to report all these values in the tables. Nevertheless, the benchmark results strengthen our case for comparing the results of $\hat{\beta}_{VA}$ across specifications.

While these results about the persistence of teacher value-added are important, there is always the possibility that the nature of value-added measures, as a sort of teacher fixed effect on

test scores, leaves them particularly vulnerable to measuring strategic teacher manipulations that add little social value, such as the ability to teach to a particular test. One possible way to explore this is to use a teacher's measure of value-added in the opposite subject (for example reading value-added as an instrument for the students past math score) as the excluded instrument. The results of this specification are reported in the second column of Table 5. While there is a statistically significant difference between this and the baseline estimates for math, the magnitude of the point estimates are virtually identical. Of course this could mean either that teaching to the test is not an important component of teacher value-added in this setting or that the ability to teach to the test has a high positive correlation across subjects.

One of the primary advantages of our method of estimating persistence is the straightforward way it provides for comparing persistence estimates across different programs or policy levers, such as alternative measures of teacher quality. Columns (3) – (4) of Table 5 demonstrate some of these comparisons using the North Carolina data. The third column presents estimates of the persistence in test score shocks associated with having a novice (first or second year) teacher. To obtain these estimates, we would like to instrument once (twice) lagged student achievement with an indicator of whether the student's once (twice) lagged teacher was a novice. However, in our data students are non-randomly assigned to teachers on the basis of various observable teacher characteristics, including experience, credentials and principal ratings. To address this concern, we purge these instruments of correlation with student characteristics by regressing the teacher characteristic in question on lagged student reading and math scores (those scores will be twice lagged relative to the final test score of interest) as well as observable student characteristics and school and year fixed effects. The residuals from these regressions serve as our instruments. This is analogous to the process of producing our teacher

value-added instruments outlined above.

In both math and reading, the coefficient estimates are significantly lower than the corresponding value-added estimates, suggesting that teacher experience shocks generate *less* persistent effects than value-added shocks. In fact, it is impossible to reject a zero persistence outcome for the effect of teacher inexperience on reading scores.

The fourth column looks at the persistence of test score shocks generated by having a teacher with National Board Certification (NBC). More specifically, we define a teacher as having NBC if he or she ever appears as certified in our data. For this reason, our measure is capturing both any teacher characteristics associated with the propensity to apply for and receive certification as well as any "effects" of the certification process itself.[18] Following the logic described above, as our instrument we use residuals from a regression of NBC status on lagged student achievement and student demographics rather than the raw NBC indicator. For math, the persistence of having a NBC teacher is roughly equivalent to having a teacher with a value-added measure one standard deviation higher than average. For reading, the point estimate of the NBC persistence is nearly zero, although the estimate is quite imprecise.

As a further specification check, we perform a similar analysis of teacher value-added and another teacher quality measure on a separate longitudinal dataset that matches students to teachers from 1998 to 2004 in an anonymous mid-sized Western school district. This allows us to check whether our persistence estimates might be sensitive to the particular set of institutions or tests given in North Carolina. Using this data in standard deviation units, we constructed analogous measures of teacher value-added and examined the persistence of test scores due to the resulting variation in teacher value-added.

The presentation of results in columns (5) – (6) shows the estimated persistence of

teacher value-added in this district is even lower than in North Carolina. The point estimates suggest a one-year persistence of 0.15 for reading and 0.17 for math.[19] In both cases, we reject the hypothesis of zero persistence. This school district also has an alternative available measure of teacher quality, a numeric principal evaluation of the teacher's performance. The persistence estimates in column 6 suggest that the components of teacher quality captured in principal evaluations have similar persistence than those captured by value-added measures, although these estimates are also quite imprecise. These results are important because they suggest that the low measured persistence of teacher value-added is not a peculiarity of North Carolina schools or one specific test metric.

While our preferred estimation strategy has a number of advantages relative to other approaches, we should be able to recover similar persistence measures using a reduced form strategy. One potential difficulty is that we do not observe actual teacher quality but rather a noisy proxy leading a regression of student achievement on raw value-added measures to exhibit substantial attenuation bias.[20] To overcome this problem, we can construct empirical Bayes measures of teacher quality, which "shrink" a teacher's estimated value-added to the grand mean of teacher quality according to the precision of the estimate. The goal is to estimate the expected teacher quality conditional upon the observed noisy signal.

Morris (1983) describes a method for implementing parametric empirical Bayes (EB) estimators.[21] Briefly, to obtain EB estimates of teacher quality, we multiply each raw value-added measure by its statistical reliability. This shrinks each estimate toward zero, which is the mean of teacher quality by construction.[22] For our reliability calculation, we obtain an estimate of the true variance by calculating the covariance between measures of teacher value-added estimated for the same teacher in different periods and calculate the estimation error by

subtracting it from the variance of teacher value-added estimates calculated with a single year of data.

To identify the reduced form impact of current teacher quality on student achievement, we estimate the following regression for fourth and fifth grade students:

(17) $test_{ijt} = \alpha\hat{\theta}_j^{EB} + \beta test_{it-1} + X_{it}\Gamma + \varepsilon_{ijt}$.

Our covariates include a student's own lagged reading and math achievement along with the demographic controls used in our baseline IV specifications. The inclusion of classroom fixed effects is not possible as it would be collinear with the current teacher's value-added score. Thus we control for classroom composition by including classroom average lagged reading and math performance. We use standard errors that are robust to clustering within the classroom. The coefficients from this regression are found in row (1) of Table 6. We see that the coefficient on the EB measure of reading and math achievement are approximately 1.1. With the EB approach, we would have expected the coefficient to be very close to one. The fact that it exceeds this value may reflect some non-random assignment of high-ability children to teachers with high observed value-added.

In row (2), we show the estimated impact of lagged teacher quality estimated from regressions that depart slightly from equation (17). We control for twice lagged achievement as lagged achievement is driven in part by lagged teacher quality. We include current classroom fixed effects which subsume the average classroom performance measures and control more effectively for classroom composition and selection. The reduced form estimates of the impact of lagged teacher quality are .20 for reading and .29 for math. These are virtually identical to our IV persistence measures. If we divide them by the initial period effects, they are slightly lower though substantively the same.

Row (3) shows the impact of twice lagged teacher value-added. Relative to the lagged specification, the only differences are that we examine students in the seventh and eighth grades and control for thrice lagged achievement. The coefficients on the twice lagged teacher value-added measures are 0.17 for reading and 0.18 for math, again very consistent with our IV persistence estimates.

VI. Conclusions

This paper develops a statistical framework to empirically assess and compare the persistence of treatment effects in education. We present a model of student learning that incorporates permanent as well as transitory learning gains, and then demonstrate that an intuitive instrumental variables estimator can recover the persistence parameter.

The primary claim of the recent teacher value-added literature is that teacher quality matters a great deal for student achievement. This claim is based on consistent findings of a large dispersion in teachers' ability to influence contemporary student test scores. While this claim may well be true relative to other policy alternatives, our results indicate that contemporaneous value-added measures are a poor indicator of long-term value-added. Indeed, test score variation due to teacher value-added is only about one-fifth as persistent as true long-run knowledge and perhaps one-third as persistent as the overall variation in test scores. Thus when measured against intuitive benchmarks, contemporary teacher value-added measures almost certainly overstate the ability of teachers, even exceptional ones, to influence the ultimate level of student knowledge.

Furthermore, when measured against other potential policy levers that involve teacher quality, value-added induced variations do not have statistically different persistence than those of principal ratings or national board certification measures. We do find, however, that value-

added variation in student achievement is significantly more persistent than the variation generated by inexperienced teachers.

Taken at face value, our results for two-year persistence imply that a policy intervention to raise teacher value-added by a standard deviation would produce a long-run effect on student math achievement closer to 0.02 standard deviations than the 0.10 standard deviation increase found in the literature (Aaronson, Barrow, and Sander 2007; Rivkin, Hanushek, and Kain 2005; Rockoff 2004).

This is likely to change how we evaluate the net benefits of programs that purport to improve, identify or retain high value-added teachers. For example, Clotfelter, Ladd, and Vigdor (2007) suggest that a teacher testing program can identify differences in teacher ability that translate into a -0.062 standard deviation effect on short-run student achievement for teachers two standard deviations below average and a 0.068 standard deviation effect for equivalently above average teachers. According to Figlio and Kenney (2007), a merit pay program to move beyond identification to retention might be expected to improve student achievement by 0.05-0.09 standard deviations. However, the cost of testing and bonuses for such a program becomes significantly harder to justify if the relevant effect size is at most 0.009-0.016 standard deviations when measured just two years later.

As mentioned earlier, our statistical model captures knowledge fadeout stemming from a variety of different sources, ranging from poor measurement of student knowledge to structural elements in the education system that lead to real knowledge depreciation. Although it is impossible in the present context to definitively label one or more explanations as verified, we can make some progress in this area. For example, our results show that the low persistence of teacher quality induced variation is not due to some flaw in the construction or use of value-

added measures, but is common to other methods of measuring teacher quality.

Our results also provide some evidence that the observed fadeout is not due to compensatory teacher assignment. Indeed, measured persistence declines when controlling for the current class-room fixed effects suggesting that teacher quality is positively correlated over time. This positive autocorrelation is observed directly when we look at the correlation between the measured value-added of the student's prior year teacher and the student's teacher two years ago.

Should the particular explanation for fadeout change how we should think about the policy possibilities of value-added? To examine this, consider under what circumstances exceptional teachers could have widespread and enduring effects in ways that belie our estimates. Three criteria would have to be met: The knowledge that students could obtain from these exceptional teachers would have to be valuable to the true long-run outcomes of interest (such as wages or future happiness), retained by the student, and not tested on future exams. To the degree that all three of these conditions exist, the implications of this analysis should be tempered.

While it is certainly possible that these conditions are all met, we believe it is unlikely that the magnitude of fadeout we observe can be completely (or even mostly) explained by these factors. For example, there are few instances in elementary school mathematics where knowledge is not cumulative. Although fourth grade exams may not include exercises designed to measure subtraction, for example, that skill is implicitly tested in problems requiring long division.

Finally, 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 methods previously used, our approach allows straightforward statistical inference and comparisons across policies. It also relates the empirical results to the assumed data generating process. This may be useful as researchers and policymakers expand their efforts to more accurately measure the long-run impact of education policies and programs.

References

Angrist, Joshua D., and Victor Lavy. 1999. "Using Maimonides Rule to Estimate the Effect of Class Size on Scholastic Achievement." *Quarterly Journal of Economics* 114(2):533-575.

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, Jishnu Das, Asim I. Khwaja, and Tristan Zajonc. 2008. "Do Value-Added Estimates Add Value? Accounting for Learning Dynamics." Working Paper 158. Cambridge: Harvard Center for International Development.

Barnett, W. Steven. 1985. "Benefit-Cost Analysis of the Perry Preschool Program and Its Policy Implications." *Educational Evaluation and Policy Analysis* 7(4): 333-342.

Carrell, Scott and J. West. 2008. "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors." Working Paper 14081. Cambridge: National Bureau of Economic Research.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2007. "How and why do teacher credentials matter for student achievement?" Working Paper 12828. Cambridge: National Bureau of Economic Research.

_____. 2006. "Teacher-Student Matching and the Assessment of Teacher Effectiveness."

Journal of Human Resources 41(4): 778-820.

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

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

Figlio, David N., and Lawrence W. Kenney. 2007. "Individual Teacher Incentives and Student Performance." *Journal of Public Economics* 91(5): 901-914.

Goldhaber, Dan, and Emily Anthony 2007. "Can Teacher Quality be Effectively Assessed? National Board Certification as a Signal of Effective Teaching." *Review of Economics and Statistics* 89(1): 134-150.

Imbens, Guido W and Angrist, Joshua D., 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62(2): 467-75.

Jacob, Brian A., and Lars Lefgren 2008. "Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education." *Journal of Labor Economics* 26(1): 101-136.

_____. 2007. "What Do Parents Value in Education? An Empirical Investigation of Parents' Revealed Preferences for Teachers." *Quarterly Journal of Economics* 122(4): 1603-1637.

_____. 2004. "Remedial Education and Student Achievement: A Regression-Discontinuity Analysis." *Review of Economics and Statistics* 86(1): 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(3): 843-77.

Kane, Thomas J., Jonah E. Rockoff, and Douglas O. Staiger. 2008. "What Does Certification

Tell Us About Teacher Effectiveness? Evidence from New York City." *Economics of Education Review* 27(6): 615-631.

Kane, Thomas J., and Douglas O. Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." Working Paper 14607. Cambridge: National Bureau of Economic Research.

_____. 2002. "The Promises and Pitfalls of Using Imprecise School Accountability Measures." *Journal of Economic Perspectives* 16(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(468): 1-28.

Lockwood, J. R., Daniel F. McCaffrey, Louis T. Mariano, and Claude Setodji. 2007. "Bayesian Methods for Scalable Multivariate Value-Added Assessment." *Journal of Educational and Behavioral Statistics* 32(2): 125-150.

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* 29(1): 67-101.

Morris, Carl N. 1983. "Parametric Empirical Bayes Inference: Theory and Applications." *Journal of the American Statistical Association* 78(381): 47-55.

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(2): 127-142

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. "Teachers, Schools, and Academic Achievement." *Econometrica* 73(2): 417–58.

Rockoff, Jonah E. 2004. "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data," *American Economic Review* 94(2): 247-252.

Rothstein, Jesse. 2009. "Teacher Quality in Educational Production: Tracking, Decay and Student Achievement." *Quarterly Journal of Economics*. Forthcoming.

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 Specification and Estimation of the Production Function for Cognitive Achievement." *Economic Journal* 113(485): 3-33.

_____. 2006. "The Production of Cognitive Achievement in Children: Home, School and Racial Test Score Gaps." Working Paper 04-019. Philadelphia: Penn Institute for Economic Research.

_____

Endnotes

1. 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.

2. 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.

3. This presupposes that understanding the concept does not facilitate the learning of a more advanced concept which is subsequently tested.

4. 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. Given a different data generating process the structural interpretation of $\hat{\beta}_{OLS}$ and $\hat{\beta}_{LR}$ may change but they will still retain the LATE interpretation as the persistence arising from all sources of achievement variation and long-run differences in achievement respectively.

6. While treat could be a binary assignment status indicator, it could also specify a continuous policy variable such as educational spending or class size.

7. In our framework, a test of the hypothesis that different educational interventions have different rates of persistence can be implemented as a standard test of over-identifying restrictions.

8. The teacher cheating in Chicago identified by Jacob and Levitt (2003) led to large observed performance increases, but was correlated to poor actual performance in the classroom. Also, Carrell and West (2008) show that short run value-added among Air Force Academy Faculty is negatively correlated to long-run value-added.

9. In general, the $\lim\left(\hat{\beta}_{VA}\right)$ is bounded between $\delta \dfrac{\sigma_{\theta_l}}{\sigma_{\theta_l} - \sigma_{\theta_s}}$ when the correlation between short and long run value-added is -1 and $\delta \dfrac{\sigma_{\theta_l}}{\sigma_{\theta_l} + \sigma_{\theta_s}}$ when it is 1.

10. We thank Jacob Vigdor for providing us with the data used in Clotfelter, Ladd, and Vigdor (2007) for our analysis.

11. The teachers identified in the student test file are those that proctored the exam, not necessarily those that taught the class. The authors describe the three-tiered system of matching

students to actual teachers. The first assigns the proctor as the teacher if the proctor taught the correct grade and subject that year. They also look at the composition of the test taking students and compare it with the composition of students in classes from the teacher file to find matches.

12. When examining two year depreciation rates we use the same students a year later when they are in grades 6 and 7.

13. Teachers who taught for only one year have a missing value-added measure. They are removed from consideration using a dummy variable for missing teacher VA in the second stage (hence their missing status is not used for identification). For comparability, we exclude single-year teachers from later reduced form estimates.

14. See, for example, Todd and Wolpin (2006) and Sass (2006).

15. The first stage estimates are very similar to the OLS persistence measures.

16. For the one year persistence estimates, the first stage coefficients on teacher value-added are 0.39 for reading and 0.69 for math. In both cases, the corresponding t-statistics exceed forty.

17. To perform this comparison, we divide teachers into terciles on the basis of their value-added. When examining the impact of being assigned a teacher in the top third, we instrument lagged value-added with a dummy variable that takes on a value of 1 if prior year teacher was in the top third of the value-added distribution. We include in the second stage a dummy variable indicating whether the prior year teacher was in the bottom third. Thus we exploit only variation due to assignment to a teacher in the top third of value-added *relative* to the middle third (the omitted category). When looking at the impact of assignment to a poor teacher, we do the opposite.

18. For a more detailed discussion of this issue, see Goldhaber and Owens (2007).

19. The benchmark estimates show an estimated persistence of student achievement captured by OLS at 0.58 on both tests and an estimated persistence of long-run knowledge of 1.07 for reading and 1.08 for math. If the slightly elevated long-run estimates are an estimate of scale, the western districts teacher value-added fadeout looks even steeper relative to North Carolina.

20. This is not a problem in the context of our instrumental variables strategy, which is robust to measurement error in teacher quality as long it is uncorrelated to the second stage residual.

21. Education researchers often use Hierarchical Linear Modeling (HLM) to obtain Bayesian estimates of teacher quality. Kane and Staiger (2008) employ an approach similar to our own and determine that the correlation between their EB estimates of teacher quality and those obtain by HLM have a correlation that exceed .99.

22. For simplicity, we have treated each classroom as if it had the same number of students. If this assumption were relaxed, the variance of the estimation error would depend on the number of students in each of a teacher's classes.

Table 1
Summary Statistics

Variable	North Carolina
Normalized reading score	0.000
	(1.000)
Normalized math score	0.000
	(1.000)
Student fraction male	0.508
	(0.500)
Student fraction free lunch	0.440
	(0.496)
Student fraction White	0.625
	(0.484)
Student fraction Black	0.296
	(0.457)
Student fraction Hispanic	0.035
	(0.184)
Student fraction special education	0.110
	(0.312)
Student fraction limited English	0.016
	(0.124)
Student age	11.798
	(0.721)
Student has novice teacher	0.123
	(0.328)
Student has board certified teacher	0.098
	(0.297)
Fifth grade	0.499
	(0.500)
Sixth grade	0.501
	(0.500)

Notes: The Table reports standard deviations in parentheses below means. Test scores

are normalized to be mean zero with unit standard deviation by state/district, year,

grade, and subject.

Table 2
Estimates of the Persistence of Achievement

	One Year			Two Year		
	$\hat{\beta}_{OLS}$	$\hat{\beta}_{LR}$	$\hat{\beta}_{VA}$	$\hat{\beta}_{OLS}$	$\hat{\beta}_{LR}$	$\hat{\beta}_{VA}$
A. Reading						
Prior year achievement coefficient	0.68** (0.001)	0.97** (0.001)	0.20** (0.02)	0.62** (0.002)	0.90** (0.002)	0.18** (0.02)
F-Statistic of instruments [p-value]	--	300,000 [0.00]	2,192 [0.00]	--	260,000 [0.00]	2,347 [0.00]
Observations	1.3 mil	1.3 mil	1.3 mil	1.1 mil	1.1 mil	1.1 mil
R-Squared	0.74	0.58	0.50	0.69	0.51	0.42
B. Math						
Prior year achievement coefficient	0.71** (0.001)	0.95** (0.001)	0.27** (0.01)	0.63** (0.001)	0.87** (0.001)	0.16** (0.008)
F-Statistic of instruments [p-value]	--	370,000 [0.00]	14,000 [0.00]	--	290,000 [0.00]	13,000 [0.00]
Observations	1.3 mil	1.3 mil	1.3 mil	1.1 mil	1.1 mil	1.1 mil
R-Squared	0.80	0.67	0.58	0.78	0.59	0.46

Notes: Reported standard errors in parentheses correct for clustering at the classroom level. **

indicates five percent significance.

Table 3
Robustness Checks

		Reading		Math	
		One Year Persistence	Two Year Persistence	One Year Persistence	Two Year Persistence
(1)	Baseline	0.20** (0.02)	0.18** (0.02)	0.27** (0.01)	0.16** (0.01)
(2)	Controlling *only* for grade, school, and year in second stage	0.54** (0.02)	0.46** (0.02)	0.49** (0.01)	0.36** (0.02)
(3)	Omitting year *t* classroom fixed effects from baseline	0.32** (0.02)	0.24** (0.02)	0.33** (0.01)	0.19** (0.01)
(4)	Value-Added using estimated achievement gains	0.10** (0.02)	0.06** (0.02)	0.25** (0.01)	0.13** (0.01)
(5)	Value-Added using estimated achievement gains normalized by initial score	0.15** (0.02)	0.12** (0.02)	0.26** (0.01)	0.14** (0.01)
(6)	Top third of teacher quality compared to middle third	0.13** (0.05)	0.03 (0.06)	0.27** (0.01)	0.12** (0.02)
(7)	Bottom third of teacher quality compared to middle third	0.25** (0.04)	0.26** (0.04)	0.29** (0.02)	0.19** (0.02)
(8)	Test performance measured in percentiles	0.20** (0.02)	0.17** (0.02)	0.28** (0.01)	0.16** (0.01)
(9)	Test performance measured in scale scores	0.20** (0.02)	0.18** (0.02)	0.28** (0.01)	0.16** (0.01)

Notes: Reported standard errors in parentheses correct for clustering at the classroom level.

** indicates five percent significance. Though not reported, estimation results for benchmark

$\hat{\beta}_{LR}$ are quite consistent across the table's specifications with estimates on always in the 0.95-

0.99 range for reading scores and the 0.93-0.97 range for math scores.

Table 4
Heterogeneity of 1-Year Depreciation Rates

Subject	Reading	Math
A. Grade		
Fifth	0.21**	0.26**
	(0.03)	(0.01)
Sixth	0.20**	0.29**
	(0.03)	(0.01)
Chi-Squared equal coefficients	0.06	4.81
[p-value]	[0.81]	[0.03]
B. Gender		
Male	0.18**	0.27**
	(0.03)	(0.01)
Female	0.24**	0.29**
	(0.03)	(0.01)
Chi-Squared equal coefficients	2.47	1.92
[p-value]	[0.11]	[0.17]
C. Year		
1998	0.14**	0.25**
	(0.06)	(0.02)
1999	0.17**	0.24**
	(0.06)	(0.02)
2000	0.25**	0.28**
	(0.05)	(0.02)
2001	0.23**	0.26**
	(0.05)	(0.02)
2002	0.27**	0.31**
	(0.05)	(0.02)
2003	0.17**	0.26**
	(0.05)	(0.02)
2004	0.20**	0.33**
	(0.06)	(0.02)
Chi-Squared equal coefficients	5.39	17.59
[p-value]	[0.49]	[0.01]
D. Race		
White	0.19**	0.26**
	(0.03)	(0.01)
Black	0.21**	0.30**
	(0.04)	(0.01)
Chi-Squared equal coefficients	0.16	5.18
[p-value]	[0.69]	[0.02]
E. Free lunch status		
Yes	0.22**	0.31**
	(0.03)	(0.01)
No	0.21**	0.26**
	(0.03)	(0.01)

Chi-Squared equal coefficients	0.02	8.41
[p-value]	[0.90]	[0.00]
F. Teacher experience		
Less than five years	0.16**	0.29**
	(0.07)	(0.02)
Five or more years	0.22**	0.28**
	(0.02)	(0.01)
Chi-Squared equal coefficients	0.59	0.14
[p-value]	[0.44]	[0.71]

Notes: Reported standard errors in parentheses correct for clustering at the classroom

level. ** indicates five percent significance. Though not reported, estimation results

for benchmark $\hat{\beta}_{LR}$ are quite consistent across the table's specifications with estimates

on always in the 0.95-0.99 range for reading scores and the 0.93-0.97 range for math

scores.

Table 5
Comparing Value-Added Persistence with Other Teacher Policies

	North Carolina				Western U.S. district	
Instrument:	Value-Added	Other Subject Value-Added	Inexperienced Teacher	National Board Certification	Value-Added	Principal Rating
	(1)	(2)	(3)	(4)	(5)	(6)
A. Reading						
Prior year achievement coefficient	0.20** (0.02)	0.23** (0.02)	0.01 (0.07)	0.01 (0.36)	0.15** (0.05)	0.22 (0.18)
F-Statistic of instruments [p-value]	2,192 [0.00]	2392 [0.00]	228 [0.00]	8 [0.01]	335 [0.00]	27 [0.00]
Observations	1.3 mil	1.3 mil	1.3 mil	1.3 mil	21,421	21,421
χ^2 equal coefficients (with value-added) [p-value]	---	1.42 [0.23]	16.19 [0.00]	0.61 [0.43]	---	0.16 [0.69]
B. Math						
Prior year achievement coefficient	0.27** (0.01)	0.25** (0.01)	0.10** (0.04)	0.22** (0.08)	0.17** (0.05)	0.15 (0.12)
F-Statistic of instruments [p-value]	14,000 [0.00]	4768 [0.00]	440 [0.00]	110 [0.00]	325 [0.00]	44 [0.00]
Observations	1.3 mil	1.3 mil	1.3 mil	1.3 mil	21,421	21,421
χ^2 equal coefficients (with value-added) [p-value]	---	8.59 [0.00]	22.29 [0.00]	0.55 [0.46]	---	0.06 [0.81]

Notes: Reported standard errors in parentheses correct for clustering at the classroom level. For the western district the benchmark

$\hat{\beta}_{LR}$ is 1.07 for reading and 1.08 for math while $\hat{\beta}_{OLS}$ is 0.58 on both tests. ** indicates five percent significance.

Table 6
Reduced Form Impact of Current and Lagged Teacher Value-Added on Student Achievement

Coefficient on Empirical Bayes Measure of Teacher Value-Added	Reading	Math
(1) Current teacher (fourth and fifth graders)	1.13** (0.02)	1.12** (0.01)
(2) Teacher one year ago (fifth and sixth graders)	0.20** (0.02)	0.29** (0.01)
(3) Teacher two years ago (sixth and seventh graders)	0.17** (0.02)	0.18** (0.01)

Notes: The reported coefficients are obtained from six separate regressions. The construction of the empirical bayes measures of teacher value-added is described in the text. In addition to the same controls used in the IV specifications, we control for prior year reading and math achievement in row (1), twice lagged achievement in row (2), and thrice lagged achievement in row (3). In rows (2) – (3), we also control for current classroom fixed effects. In row (1), the classroom fixed effect is collinear with current teacher value-added so we control for control for classroom composition with the average prior year reading and math scores. Reported standard errors in parentheses correct for clustering at the classroom level. ** indicates five percent significance.