Estimating Teacher Value-Added in a Cumulative Production Function [*]

Josh Kinsler, University of Rochester

Preliminary - Please Do Not Cite

October 28, 2008

Abstract

Recent teacher value-added studies find that teachers play a significant role in the production of student achievement. However, much of this work makes unattractive assumptions about the persistence of teacher inputs, that if incorrect, will bias not only the importance of teacher quality in the production process, but will also cloud any interpretation regarding the long-run impact of teachers. I develop and estimate a cumulative production function that explicitly accounts for the accumulation of past teacher inputs. I find that teachers play a larger role in contemporaneous outcomes than previously believed, but that their effect on student achievement is rather short-lived.

1 Introduction

Quantifying the impact of schools and teachers in the production of human capital has long been of interest to researchers, school administrators, and parents. Education researchers and economists hope to increase efficiency and improve output by identifying the influence of various factors in production, while parents and administrators simply seek the ability to identify excellent schools and teachers. The seminal work in this line of research is the famed Coleman Report(1966), which found that the effect of schools and teachers was easily swamped by family background and peer inputs. Coleman's findings clashed with widely held beliefs

[*] I would like to thank the applied lunch group at the University of Rochester.

regarding the role of schools and teachers in student development and spawned an enormous literature seeking to reconcile these differences.

While early findings tended to support the initial result that teachers had little effect on student performance[1], more recently researchers have reached a consensus that teachers do in fact play a significant role in producing student achievement.[2] This evolution largely tracks advancements in data collection and computing power. Initial attempts to evaluate the impact of teachers used cross-sectional data to relate observed teacher characteristics to student performance. The results suggested that besides experience in the first few years, most observable teacher characteristics had little impact.[3] This led to the idea that much of what makes a teacher effective is unobserved. Estimating unobserved teacher quality has only recently been possible as a result of increased computing power and the development of large institutional education data linking student outcomes to teachers over multiple years and cohorts. Three papers that exemplify this strategy are Rockoff (2004), Hanushek et al. (2005), and Aaronson et al. (2007), who despite varying econometric specifications find that a one standard deviation increase in teacher quality yields approximately 0.1% of a standard deviation increase in math test scores and slightly smaller effects in reading.

Identifying unobserved teacher effectiveness ex-post may not be useful for improving hiring practices, however it can be used as the basis for accountability or compensation. Many states and school districts currently uses some version of a value-added model to evaluate teachers, though there remains significant opposition to these programs.[4] The resistance to tying compensation or promotion to teacher value-added estimates stems partly from the widespread belief that the current models used to identify teacher effects are inadequate. Two oft-cited criticisms of these models are that they fail to account for the sorting of students and teachers based on unobservable student characteristics and that they make unreasonable assumptions about the decay of past inputs.[5] Both issues result in biased estimates of teacher quality that

[1] See Boardman et al. (1979) for example.

[2] Hanushek et al. (2005), Rockoff (2004), Aaronson et al. (2007)

[3] See Goldhaber and Brewer (2000) and Clotfelter et al. (2007) for more recent applications.

[4] For example, Tennessee has been using a value-added model to inform administrators on the performance of schools since 1992. In October of 2008, New York City schools announced they would begin generating teacher data reports from a value added model framework. However, in order to appease the teachers union the city had to agree not to use the reports to make any tenure or salary decisions.

[5] See Clotfelter et al. (2006) for evidence regarding the sorting of students to teachers. Todd and Wolpin

could go in either direction depending on the true sorting method and the assumptions about the rate of input decay.

A handful of studies have attempted to tackle the issue of non-random sorting of students and teachers. Hanushek et al. (2005) collapse student data to the grade level and then difference across cohorts to remove the influence of systematic sorting among teachers within a school grade. As a result they are unable to estimate the variance of teacher quality within a grade and are unable to estimate the full distribution of teacher quality. Rockoff (2004) models the level of student achievement as a function of unobserved student and teacher effects. However, this only eliminates student sorting based on permanent differences in unobserved characteristics that affect the level of achievement. Any permanent heterogeneity in the growth of achievement will continue to bias the teacher estimates.

While significant progress has been made on the issue of student sorting, little advancement has been made in estimating cumulative models of learning that account for teacher inputs. Rather than deal directly with the accumulation of knowledge, researchers have generally made one of three rather unattractive assumptions about the persistence of past inputs: no decay, full decay, or identical geometric rates of decay for all inputs, including unobserved ability.[6] Any of these three assumptions generate a simplified model of student achievement that can be estimated controlling only for the contemporaneous teacher. However, if the assumption about the decay rate is incorrect, the individual estimates of teacher quality and the variance of teacher quality will be biased, as illustrated by Rivkin (2006) and Todd and Wolpin (2006).

Perhaps even more important is how the assumption regarding the persistence of inputs influences our interpretation of teacher quality and its long term impact on student achievement. As Jacob et al. (2007) points out, if teacher inputs decay rapidly, having an exceptional teacher today will not alter the level of human capital ultimately attained. Thus policies focused on boosting teacher value-added may yield disappointing educational returns in the long run. Using an ingenious instrumental variables strategy, Jacob et al. (2007) estimates a decay rate of teacher value added equal to about 0.2. However, in order to generate his instruments, a measure of teacher value added must first be estimated from a specification that suffers from both the sorting and decay issues.

(2006) and Rivkin (2006) illustrates the pitfalls associated with incorrect assumptions regarding decay.

[6] Rockoff (2004) assumes full decay while Aaronson et al. (2007) assumes an identical geometric rate of decay.

In this paper I develop a cumulative model of student learning that accounts for both the sorting of students and teachers based on unobservable student attributes and directly estimates the rate of decay of teacher value added. The assumptions necessary to arrive at consistent estimates of teacher effects and decay rates are no more onerous than the typical value added assumptions. The innovation is the ability to estimate large dimensional vectors of fixed effects and interactions between these fixed effects and other model parameters in a reasonable amount of time. The estimation strategy extends from the framework outlined in Arcidiacono et al. (2006).

I add to the initial cumulative model of student achievement features such as heterogeneity in the discount rate, contemporaneous and persistent classroom inputs besides the teacher, and time-varying teacher attributes. These extensions allow for the identification of critical periods in the production of achievement, as well as the distribution of teacher quality across the experience spectrum. The estimation procedure become slightly more complicated in the expanded framework, but remains quite manageable.

Using student data from North Carolina's elementary schools, I find that teachers do in fact play a large role in the production process. A 1 standard deviation increase in teacher quality is equivalent to 0.2% of a standard deviation in the level of student test scores. The fact that this result is larger than previous estimates is not surprising since most other value-added models are estimated with test score levels assuming full decay of previous inputs. If the decay rate is not zero the variance in teacher quality will be biased downward. While there is significant variation in teacher quality, teacher value-added decays at very fast rate, approximately equal to 0.22.

These seemingly contradictory results can be interpreted in one of two ways. First, teacher value-added estimates are a poor metric with which to rate teachers by since they do not reflect any long-term effect. As a result, providing incentives for teachers to boost their value-added may reallocate effort away from more productive activities, such as promoting a desire to learn or teaching social and behavioral skills. A second interpretation is simply that the exams are designed in such a way as to provide a clean signal of student ability in each grade. Thus, teachers from previous grades will have little effect since their inputs are much less relevant. In this scenario, if the tests themselves measure market relevant skills, than incentives to increase teacher value-added may be beneficial. Differentiating between these hypotheses is left for future research.

A number of other interesting results are obtained from the model. Within schools, it appears that students with low unobserved ability are assigned to higher quality teachers. This indicates that models that do not explicitly control for the non-random matching between students and teachers will likely understate the role of teachers in the production process. Teacher performance improves significantly with increases in experience. However, more experienced teachers tend to have lower unobserved ability.

The remainder of the paper is as follows. Section 2 outlines the education production function, discusses identification of the key parameters, and illustrates the estimation strategy. Section 3 introduces the North Carolina student data used to estimate the cumulative production function. Section 4 contains analysis of the model results and Section 5 concludes.

2 Cumulative Model of Education Production

2.1 Baseline Framework

Student achievement in grade g reflects the cumulative impact of student, family, and school inputs. As the focus of this paper is on the role of teachers in the educational process, I assume for the moment that teachers are the only school input in the production of achievement. Teacher input is captured through a set of unobserved fixed effects, reflecting the fact that observable teacher characteristics often have little explanatory power in predicting student test scores. Achievement for student i in grade g is given by the following general formula:

$$A_{ig} = f_g(T_i(g), \alpha_i, \epsilon_i(g)) \tag{1}$$

where A_{ig} is the achievement score of student i in grade g, $T_i(g)$ contains the full history of teacher inputs through grade g, α_i contains all non-school inputs (both family and individual) that do not vary over time, and $\epsilon_i(g)$ contains all time-varying non-school inputs and measurement error in any year. If f_g is linear and past inputs decay at a rate equal to $(1-\delta)$, achievement in grade g takes the following form:

$$A_{ig} = T_{ig} + \tau_g \alpha_i + \epsilon_{ig} + \sum_{h=1}^{g-1} (1-\delta)^{g-h} T_{ih} \tag{2}$$

This is the baseline education production function I am interested in estimating.

While the key innovation in the baseline technology is the ability to handle both the non-random sorting of students with teachers and the cumulative effect of teachers, I want to briefly acknowledge the implications of the linearity assumption. Contemporaneous complementarity between teacher and non-school inputs and complementarity between teacher inputs across grades is ruled out. In other words, the achievement contribution of a grade g teacher is homogenous with respect to non-school inputs, including innate student ability, and the individual patterns of teachers prior to grade g. This is a common assumption in value-added models, though it is possible to allow for varying teacher effects based on interactions between observable student and teacher characteristics such as gender or race.

Also common in the recent teacher value-added literature is the approach used here for dealing with the non-random sorting of students with teachers. Rather then include a host of observable student and family background characteristics, both observed and unobserved non-school inputs are captured through a permanent component, α_i. The identification argument is that conditional on α_i, teacher assignments are random. This allows for consistent estimation of the T_{ig} in the presence of non-random sorting. There are an number of assumptions that are implicit in this identification argument.

The first assumption underlying this argument is that home inputs, such as parent involvement, do not respond to the assignment of T_{ig}.[7] Given the lack of information on home inputs in most state administrative data, this assumption is difficult to avoid. The bias induced by parental responses could plausibly bias downward or upward the variance of teacher quality. If parents' effort increases (decreases) when their child is assigned a poor (good) teacher, then teacher-value added estimates will be biased towards zero. Strong complementarity between home and school inputs would lead to the opposite pattern in parental responses and lead to upward biased estimates of the importance of teacher quality.

Although the assumption of time-invariant home inputs is difficult to test using state administrative data, a second assumption underlying the identification argument, conditional

[7] Todd and Wolpin (2006), using data from the National Longitudinal Survey of Youth 1979 Child Sample (NLSY79-CS), consistently reject exogeneity of family input measures at a 90 percent confidence level, but not at a 95 percent confidence level. However, they have very limited measures of school inputs and the coefficients on these inputs are statistically insignificant whether home inputs are exogenous or endogenous. Thus it is difficult to gauge how parents might respond to individual teacher assignments.

strict exogeneity, can be tested. As Rothstein (2008) points out, because α_i is estimated using the mean achievement level (or gain), all the time-varying non-school inputs enter into the achievement equation in each period. As a result, consistent estimation requires that the transitory error in one grade and the teacher assignment in that grade or any other grade must be uncorrelated. Using a direct application of Chamberlain's correlated random effects model, Rothstein (2008) tests and rejects the strict exogeneity assumption using administrative data from North Carolina. However, Rothstein (2008) uses an asymptotic approximation to generate his test statistic, when in practice the estimating sample is quite small. As a result, the size of the proposed test is significantly distorted, as shown in Kinsler(2008). Using a similarly constructed data set with an appropriately sized test results in a failure to reject the assumption of conditional strict exogeneity.

In addition to the assumptions regarding conditional strict exogeneity and time-invariant home inputs, one additional assumption is necessary in order to consistently estimate T_{ig}. This assumption states that the number of teachers is held fixed as the sample size grows. If this were not the case, the sampling errors would not converge to zero as the sample size grows toward infinity. One way to satisfy this assumption is to grow the sample by adding cohorts of students where the teachers are held fixed over time. In estimation I use multiple cohorts of students to identify the teacher effects, largely satisfying this restriction.[8]

While the three assumptions outlined above are in concert with previous research on teacher value-added, it is the absence of an assumption about how past inputs decay that differentiates this production technology from previous models. In order to account for the accumulation of knowledge, the typical value-added analysis assumes that teacher inputs either decay entirely ($\delta = 1$) or not at all ($\delta = 0$), neither of which seem appealing. The reason for fixing δ is that it makes estimation rather simple. The two equations below show how the baseline production technology in Equation (2) simplifies when it is assumed that $\delta = 1$ or $\delta = 0$,

$$A_{ig} = T_{ig} + \tau_g \alpha_i + \epsilon_{ig} \qquad (\delta = 1) \tag{3}$$

$$A_{ig} - A_{i(g-1)} = T_{ig} + (\tau_g - \tau_{g-1})\alpha_i + e_{ig} \qquad (\delta = 0) \tag{4}$$

[8] In practice, teachers come in and out of the sample over time, such that there are some teachers observed only once. As a result, it will be necessary to correct for the sampling error in the estimated teacher effects.

where e_{ig} can be thought of as a shock to to student growth rather than the level of achievement. If τ_g equals some constant c for all grades g, then teacher value-added can be estimated as a levels equation, similar to the work by Rockoff (2004), or as a growth equation where the unobserved student heterogeneity cancels out. On the other hand, if $\tau_g - \tau_{g-1}$ is equal to some constant c for all grades g, teacher value-added can be estimated from a growth equation where student heterogeneity in the growth rate of achievement is accounted for, such as in Harris and Sass (2006a).[9]

The attractiveness of the models outlined above are their tractability, but incorrect assumptions about the decay rate will lead to biased estimates of teacher quality. Rivkin (2006) shows that when teacher or school inputs are measured by an observable characteristic, the analytical bias for the parameter on the observable input will be proportional to $\frac{\delta-1}{2}$ or $\frac{\delta}{2}$ when it is incorrectly assumed that inputs decay fully or not at all. A similar bias will exist when the input is unobserved, however, now the measure of interest is the estimated variance of the unobserved input. To illustrate the bias I perform some simple monte carlo exercises.

I generate test score data assuming a decay rate of 0.25, 0.5, or 0.75 for past teacher inputs, while estimation assumes either no decay or full decay. Student test scores in grade g are constructed according to the following simple formula,

$$A_{ig} = \tau_g \alpha_i + T_{ig} + \sum_{h=2}^{g-1} (1 - \delta_{TRUE})^{g-1-h} T_{ih} \tag{5}$$

where α_i is student ability and T_{ig} is the value-added effect of student i's teacher in grade g. Each student is observed four times and their are four cohorts of students. Classes consist of 15 students and each school has four classes per grade. I assume a baseline score exists for each student that is only a function of α_i. I also assume $\tau_g - \tau_{g-1} = 1$, resulting in student heterogeneity in the growth of scores. As a result, I estimate the model on the growth in student achievement as opposed to the level of achievement. Notice however, that Equation

[9] One other approach for dealing with the decay of past inputs is to regress test scores in grade g on all grade g inputs and the test score from grade $g-1$. The assumption underlying this approach is that all inputs from grades $g' < g$ decay at the exact same geometric rate. This implies that the effect of the teacher in grade $g-1$ decays at the same rate as say unobserved ability. Even if this assumption were accurate, as Harris and Sass (2006b) points out, OLS estimation will still yield biased results since the lagged test score will be correlated with the new composite error term.

(5) contains no sampling error. Thus, any bias in the estimate of the variance of teacher quality must arise from the model being incorrectly specified.

Table 1 shows that an assumption of no decay or full decay lead to significantly biased estimates of the variance of teacher quality. When the true decay rate is 0.5 and it is assumed that past teacher inputs do not decay at all, the estimated variance of teacher quality is biased upwards by approximately 35%. The magnitude of the bias when past inputs are assumed to decay completely is similar at around 40%, but with the opposite sign. The upward bias occurs since in the case of no decay, past teachers drop out of the growth score equation. Thus the contemporaneous teacher estimate must account for the missing variance from previous teachers. The opposite occurs in the case when no decay is assumed. Now both the contemporaneous teacher and past teachers have equal weight in the growth equation. The variance of both teacher estimates must be scaled down in order to account for the fact that the past inputs should be scaled by the decay rate. Not surprisingly, when the true decay rate is 0.25(0.75), the full(no) decay model performs significantly better.

Since incorrect assumptions about the decay of teacher inputs clearly lead to incorrect teacher value-added estimates, I propose to let the data speak to how quickly teacher inputs decay. The variation in the data that allows for the identification of δ comes from the re-sorting of students across classrooms between grades. As an example, assume teacher's A and B teach 3rd and 4th grade respectively. Conditional on being in Teacher B's 4th grade class, there must be variation in the 3rd grade teacher, otherwise δ could not be pinned down (nor could the teacher value-added for A and B unless δ was assumed to equal 1). Because multiple cohorts of students will be used in estimation, perfect tracking within cohorts is allowed, as long as the string of teacher assignments is not the same across cohorts. For example, now assume teachers A and B teach 3rd grade and teachers C and D teach 4th grade. In cohort 1 all of teacher A's students are matched with teacher C, and all of teacher B's students are matched with teacher D. Perfect tracking in the second cohort is allowed as along as all of A's students are matched with D in 4th grade.

2.2 Estimation

In contrast to the rather straightforward argument for the identification of δ, estimating δ is somewhat more complicated, precisely the reason why it is often assumed to equal zero or one.

The difficulty in estimating δ in this framework is that it is interacted with an unobserved fixed component that must also be estimated. Rather than work in levels, I first difference the achievement outcomes, eliminating any unobserved student heterogeneity in the level of achievement. The growth in achievement from grade $g-1$ to g is then given by:

$$A_{ig} - A_{i(g-1)} \quad = \quad T_{ig} + (\tau_g - \tau_{g-1})\alpha_i - \delta \sum_{h=1}^{g-1} (1-\delta)^{g-1-h} T_{ih} + e_{ig} \tag{6}$$

where for tractability purposes I assume that $\tau_g - \tau_{g-1}$ is equal to one for all grades g.[10] The inclusion of α_i in the growth model allows for student heterogeneity in the trajectory of achievement.

The goal is then to find the solution to the non-linear least squares problem given by

$$\min_{\alpha,\delta,T} \sum_{i=1}^{N} \sum_{g=1}^{G} \left(\tilde{A}_{ig} - \alpha_i - T_{ig} - \delta \sum_{h=1}^{g-1} (1-\delta)^{g-1-h} T_{ih} \right)^2 \tag{7}$$

where $\tilde{A}_{ig}$ is the growth in student achievement from grade $g-1$ to g. Minimizing the objective function with respect to α, δ, and T in one step is infeasible as there will be thousands of students and teachers in the data. Rather, I take an iterative approach that extends the basic estimation strategy outlined in Arcidiacono et al. (2006) to allow for the accumulation of teacher inputs over time. The iterative strategy toggles between estimating (or updating) each parameter vector taking the other sets of parameters as given. Because the sum of squared errors is decreased at each step, the process will eventually converge to the set of parameter values that minimizes the least squares problem in Equation (7).

In practice, the estimation procedure follows the steps listed below. The algorithm begins with an initial guess for α and T and then iterates on three steps, with the qth iteration given by:

- Step 1: Conditional on α^{q-1} and T^{q-1}, estimate δ^q by non-linear least squares.
- Step 2: Conditional on δ^q and T^{q-1}, update α_i^q by setting the derivative of the least squares problem with respect to α_i equal to zero. Solving for α_i yields

$$\alpha_i^q = \frac{1}{G} \sum_{g=1}^{G} \left(\tilde{A}_{ig} - T_{ig} - \delta \sum_{h=1}^{g-1} (1-\delta)^{g-1-h} T_{ih} \right) \tag{8}$$

[10] In practice it is possible to estimate $\tilde{\tau}_g = \tau_g - \tau_{g-1}$, normalizing $\tilde{\tau}_2 = 1$. However, because the student fixed effects are not estimated consistently, $\tilde{\tau}_g$ would be inconsistent (standard incidental parameters problem in a non-linear model) and possibly contaminate the teacher value-added estimates and the estimate of δ.

- Step 3: Conditional on α^q and δ^q, update T_j^q (where j indexes a particular teacher) by setting the derivative of the least squares problem with respect to T_j equal to zero and solving for T_j, as shown in Equation (9).

The third step in the iterative procedure is significantly more complicated than the first two. First, the derivative of the least squares problem with respect to T_j, the teacher-value added for teacher j, is a complicated function that depends on the grade level of the teacher. For example, a teacher in grade G has no measurable long term effect since there are no future achievement outcomes beyond grade G, and thus the derivative is rather straightforward. A teacher in grade 1 will have a lasting impact on each student through grade G, and this impact is used to accurately estimate a teacher's true effect.

Taking the derivative of the least squares problem with respect to T_j, where teacher j teaches students in grade g, yields

$$T_j = \frac{\sum_{i=1}^{N_j}\left[(\tilde{A}_{ig} - \alpha_i - \delta\sum_{h=1}^{g-1}(1-\delta)^{g-1-h}T_{ih}) - \delta\sum_{h=g+1}^{G}\left((1-\delta)^{h-g-1}\Big(\tilde{A}_{ih} - \alpha_i - T_{ih} - \delta\sum_{\substack{k=1\\k\neq g}}^{h-1}(1-\delta)^{h-1-k}T_{ik}\Big)\right)\right]}{\sum_{i=1}^{N_j}\left[1 + \delta^2\sum_{h=g+1}^{G}(1-\delta)^{2(h-g-1)}\right]} \tag{9}$$

where N_j is the total number of students taught by teacher j in grade g across all cohorts. In the numerator, the first term in parentheses captures the contemporaneous impact of teacher j in grade g. It is simply the residual of the grade g outcomes for each student i, net of T_j. The second term in the numerator, consisting of the sum from grade $g+1$ to G captures the lasting impact of teacher T_j. Again, it is simply the residual of the grade $g' > g$ outcome net of the discounted contribution of teacher T_j. To evaluate Equation (9), α_i and δ take their qth iteration value, while T_k for $k \neq j$ are evaluated at their $q-1$ iteration value.

Not only are we interested in the extent to which inputs persist, but also the importance of teachers in the production function relative to other inputs. This is captured by the variance of the estimated teacher effects. However, the variance of $\hat{T}$ will yield a biased estimate of the true variance in teacher quality since it will also contain any sampling variance in the teacher quality estimates. To correct for this I bootstrap the above procedure, drawing with replacement from the estimation sample. Using the estimates of teacher quality across the bootstrap samples I construct a measure of the sampling variance for each teacher. I subtract the average of these values from the variance of $\hat{T}$ to arrive at an estimate of the true variance in teacher quality.

2.3 Extensions

The baseline achievement production technology outlined in the previous section is restrictive along a number of dimensions. This section outlines some simple extensions to the basic model that allow for heterogeneity in the discount rate, other classroom inputs, and variation in teacher quality.

2.3.1 Heterogeneity in δ

The discussion in the previous section relied on the assumption of a constant geometric rate of discount. In reality, the knowledge imparted at a certain age may matter more for future performance than inputs in other years. If this is true, it would suggest that some grades may be more critical than others and that the assignment of teachers should account for this.[11] The baseline production function can easily accommodate this by simply indexing δ by grade, as seen below.

$$A_{ig} = T_{ig} + \tau_g \alpha_i + \sum_{h=1}^{g-1} (1-\delta_h)^{g-h} T_{ih} + e_{ig} \tag{10}$$

The identification argument is the same, except now it is critical that students do not return to their same class configurations two or three years into the future. The steps necessary for estimation remain largely the same, except that in Step 1, I estimate multiple δ's by non-linear least squares. Also, in the first order conditions for α_i and T_j, the δ's will be indexed by grade.

In addition to relaxing the homogeneity of δ, it is also possible to relax the assumption that inputs decay at a geometric rate. Teacher inputs may decay very quickly after one year, but then reach a steady state where the effects no longer decay. This would imply that teachers early in the education process have a significant long term effect on achievement growth. Again, schools could use this information to find the optimal teacher allocation. The

[11] This is similar to the idea of critical periods discussed in.

production function would now take the following form

$$A_{ig} = T_{ig} + \tau_g \alpha_i + \sum_{h=1}^{g-1} (1 - \delta_{g-h}) T_{ih} + e_{ig} \tag{11}$$

where δ is indexed according to how many periods have passed since the input was applied. Again, the estimation procedure is altered to account for the multiple discount rates.

The two extensions discussed in this section also help differentiate this approach for estimating teacher-value added from previous approaches that rely on lag scores, such as Aaronson et al. (2007). In a lag score framework, not only is it critical that all inputs decay at the same rate, including teacher, school, and student inputs, but that the decay rate be homogenous and geometric. If this is not the case, then the standard simplification in which all past inputs drop out of the levels equation no longer holds, essentially invalidating this approach.[12]

2.3.2 Other Classroom Inputs

In addition to the teacher, other classroom inputs can have a significant impact on student performance. Variables like class size, gender composition, and racial composition can alter the learning environment in any particular class. If excluded from the production function, these omitted variables will bias the teacher value-added estimates when classroom level attributes are correlated with both teacher ability and student performance. As an example, class size is generally believed to negatively impact student performance. If principals assign their best teachers the largest classes, the teacher value-added estimates will be biased downward.

To account for these observable classroom attributes, I extend the baseline production function in the following manner

$$A_{ig} = \beta X_{ig} + T_{ig} + \tau_g \alpha_i + \sum_{h=1}^{g-1} (1 - \delta)^{g-h} T_{ih} + e_{ig} \tag{12}$$

where X_{ig} incorporates the class attributes experienced by student i in grade g.[13] Notice that it is critical here to have multiple cohorts with which to estimate the model. With only one cohort it would not be possible to separate the classroom attributes from the teacher effect. Extending the estimation procedure to allow for the classroom environment to have

[12] See Harris and Sass (2006a) and Todd and Wolpin (2006) for further discussion of the lag score model

[13] Implicit here is an assumption that classroom characteristics affect achievement in the same manner across grades. This restriction can also be relaxed.

a contemporaneous impact is quite simple. Again, Step 1 simply needs to be extended to estimate both β and δ by non-linear least squares. In the remaining two steps, $X\hat{\beta}$ just becomes another component of the residual when updating.

Of course, allowing non-teacher inputs to have only a contemporaneous effect imposes the strong assumption that classroom inputs decay entirely from one year to the next. Given that this is not consistent with a cumulative production process, one option is to assume that classroom inputs and teachers decay at the same rate. This assumption is not particularly appealing since teachers impart specific skills while classroom attributes are likely much more transitory. Within the cumulative production technology it is possible to separately estimate the decay rates associated with the two types of input. Under this assumption, achievement for student i in grade g is now

$$A_{ig} = \beta X_{ig} + T_{ig} + \tau_g \alpha_i + \sum_{h=1}^{g-1} \Big((1-\gamma)^{g-h}(\beta X_{ih}) + (1-\delta)^{g-h}(T_{ih}) \Big) + e_{ig} \qquad (13)$$

where γ is the decay rate associated with classroom observable characteristics. Estimating this model is a simple extension of the one highlighted in the previous paragraph, where now β, γ, and δ are all estimated in Step 1.

2.3.3 Varying Teacher Quality

With one cohort of students, it is logical to assume that a teacher's effectiveness is fixed. However, if the model is to be estimated using multiple cohorts of students, assuming that teacher effectiveness is constant over multiple years conflicts with previous research. Teacher experience is one of the few observable characteristics that appear to influence student performance. Thus, we would expect teacher effectiveness to improve across multiple cohorts, at least for the teachers with the fewest years of experience.

The achievement production function can easily accommodate changes to teacher effectiveness over time. The previous section shows how it is possible to add observable classroom characteristics. One of these classroom characteristics could be teacher experience, or teacher education level. Just as with the other classroom characteristics, the effect of these teacher characteristics can decay over time. A nice feature is that the model can allow for the classroom characteristics to decay at one rate, while the composite teacher effect (innate ability plus experience, etc.) can decay at a separate rate. The interpretation of the teacher value-added coefficient will then be the expected long-term teacher effectiveness once sufficient experience is obtained. When evaluating an individual teacher, this would seem to be exactly what a principal is interested in. The achievement equation with time-varying teacher characteristics takes the following form.

$$A_{ig} = \beta_1 X_{ig} + \beta_2 X_{T_{ig}} + T_{ig} + \tau_g \alpha_i + \sum_{h=1}^{g-1} \Big((1-\gamma)^{g-h}(\beta_1 X_{ih}) + (1-\delta)^{g-h}(\beta_2 X_{T_{ih}} + T_{ih}) \Big) + e_{ig} \tag{14}$$

Expanding the first step of the iterative procedure to estimate β_2 along with slight changes to the first-order conditions for α_i and T_j are all that is necessary to estimate a gains model based on Equation (14).

Finally, an implicit assumption in all of the above specifications is that a grade g teacher's effect in grade $g' > g$ is always proportional to the grade g effect. In other words, there are no components of the teacher input that are useful only in grade g. As an example, one criticism of high-stakes testing is that teachers have an incentive to simply teach to the test. In the framework outlined above, the test specific skills learned in grade $g < g'$ will also be useful in grade g'. If the education production function is estimated using state administrative data, this assumption seems quite logical. Knowledge tested in grade g will generally build directly from knowledge tested in grade $g-1$, and so on. This assumption would of course be violated if a teacher cheated by changing student answers or sharing specific test questions prior to the exam.

2.4 Monte Carlo Evidence

To provide some small sample evidence regarding the performance of the baseline estimator and the extensions outlined in the previous sections I conduct some monte carlo experiments. For each scenario the underlying structure of the data is identical. I generate four cohorts of students, each cohort containing 4000 students. These students are randomly assigned to 50 schools. Four student outcomes are observed, the first of which is not associated with a teacher. The following three outcomes are generated from classroom assignments that may either be random or sorted by student and teacher quality. Classes consist of twenty students each and teachers are observed with four different classes of students. There is student level heterogeneity in the growth of student test scores, and as a result I estimate the models using three growth scores for each student. The structure of the generated data is quite similar to the form of the actual data I use to estimate the model.

Prior to discussing the results of the monte carlo experiments, I want to briefly describe how the variance of teacher quality is calculated. There are two differences between the monte carlo exercises here and the earlier exercises illustrating the bias associated with model misspecification. First unobserved shocks to student achievement are included in the growth scores. As a result, the fit of each model is typically around 0.65. Because of the sampling error, the estimation error in the teacher fixed effect estimates must be accounted for when calculating the variance of teacher quality. I follow the procedure described in the estimation section to arrive at the estimates of the variance in teacher quality. Second, teacher quality is now only identified within a school-grade combination. As a result, I normalize the average teacher quality within each school-grade combination to zero and account for these normalizations when calculating the overall variance in quality.

Table 2 lists the results of the monte carlo exercises. The final two columns of Table 2 illustrate the accuracy with which the overall variance in teacher quality is estimated. Across the various specifications, the variance in teacher quality is overstated by an average of just 1%. The estimates of the other parameters, such as the decay rate on past inputs, are quite accurate. The true parameter values, listed at the top of each column, are well within the 95% confidence intervals of the estimates. In the models with multiple γ's the standard errors do get large. Increasing the number and size of each cohort will result in more precise estimates. For the models containing classroom and teacher attributes, I assign each class an observable characteristic drawn from a uniform discrete distribution on the interval $[15, 30]$. This proxies for class size. Each teacher is assigned an experience level for the first cohort that is incremented by one for each successive cohort. The parameter estimates for this model are quite precise. Overall the various flavors of the model perform quite well in estimating both teacher value added and the production function parameters governing the accumulation of knowledge.

3 Data

I use administrative data on public school students in North Carolina made available by the North Carolina Education Research Data Center to estimate the model. The data contain the universe of public school students, teachers, and schools across the state. Included in the data are observable attributes of the students, including test scores, and observable attributes of teachers, such as experience. The following paragraphs describe the steps taken to refine the data.

Since I am interested in the impact of teachers, the ability to link student outcomes with individual teachers is imperative. Therefore, I only include students in grades 3-5 who are in self-contained classrooms. The teachers for these classes can always be identified, however, teacher characteristics are not available for all of these teachers.[14] Thus I create two estimation samples, one that includes only matched teachers and another that includes all teachers. The advantage of including all teachers is a much larger sample size, while utilizing the matched sample allows me to estimate models allowing for teacher quality to vary over time through experience. In addition, I drop any school that offers only one third, fourth, or fifth grade class in any year. Identification of the teacher value-added is not possible in these schools since there is no switching across grades.

While math test scores have been collected since 1995, I focus on four cohorts of students who began second grade between 2001 and 2004. I eliminate the earlier cohorts since the variance of the test score distribution changed dramatically between 2000 and 2001. A nice feature of the data however is that at the beginning of third grade, students are administered a pre-test. I use this pre-test as an initial student observation that is unaffected by teacher inputs. All past inputs, including home and school inputs will then be absorbed into the estimated student fixed effect. All test scores are normalized such that the pre-test score for the first cohort is distributed with mean zero and variance equal to one.

Besides these two primary restrictions, there are a number of other criteria I use to limit the sample. Because I am estimating a cumulative model it is imperative that I have a complete panel for each student. Thus I include only students observed each year. Finally, to ensure

[14] If the teacher cannot be matched then it is not possible to determine if they taught a self-contained class. However, if all the other teachers in that school teach self-contained classes then I assume that the non-matched teachers also teaches a self-contained class.

that the teacher value-added estimates do not rely on just one or two student observations, I eliminate any teacher with fewer than five student observations. These final two steps are completed iteratively since the elimination of a teacher will yield an unbalanced panel for some students. Eliminating these students can then lead to too few observations per teacher and so on.

Table 3 compares the population of elementary school students for the included cohorts to the two estimation samples generated according to the above criteria. Not surprisingly the two estimation samples are significantly smaller and include students with higher test scores than average. The higher test scores in the restricted samples reflect the balanced panel requirement, since students who remain on track and do not move in and out of the data tend to score better. However, the growth in test scores across the three samples is nearly identical. With regards to the teacher data, the estimation samples do yield fewer student observations per teacher. However, on average each teacher is observed with thirty students and at least two separate cohorts. For the matched sample, the average years of experience is 14.

4 Results

In this section I present the results of the cumulative model of production using the North Carolina public school data. There is one additional variable added to the model not discussed previously. The average growth in test scores from one grade to the next is not equal across grades. Rather than assuming this reflects differences in teacher quality across grades, I assume that this reflects curriculum or testing differences across grades. Thus in the growth equations I include a grade fixed effect. The baseline specification controlling only for student heterogeneity, teacher heterogeneity, and the persistence of past teacher inputs takes the following form

$$\tilde{A}_{ig} = \alpha_i + T_{ig} + \delta \sum_{h=1}^{g-1} (1-\delta)^{g-1-h} T_{ih} + \eta_g + e_{ig} \tag{15}$$

where $\tilde{A}_{ig}$ is the growth in test scores from grade $g-1$ to g and η_g is the average growth in test scores from $g-1$ to g.

4.1 Baseline

I begin by comparing the results obtained when estimating the above equation with the results obtained from a simpler growth model that assumes $\delta = 0$. When $\delta = 0$, first-differencing grades yields an outcome equation that depends only on the contemporaneous teacher. Estimation in this case is still not straightforward since the model still contains both student and teacher fixed effects. Taking the same iterative approach outlined earlier, estimation is relatively straightforward. All the results in this and the next section utilize the base sample outlined in Table 3.

Table 4 contrasts the results across the two specifications. The first thing that jumps out is the extremely high rate at which teacher inputs decay. $(1-\delta)$ is equal to 0.233 and precisely estimated. This essentially means that by the time three years have passed, any value-added impact of a teacher has disappeared. There are two ways to interpret this result. First, teacher value-added is a poor metric with which to evaluate teachers since it is not indicative of any long-term gain in student achievement. Thus policies that emphasize improving teacher value-added may induce teachers to decrease effort on other unobserved metrics that have a greater permanent effect. The second interpretation is that the curriculum and tests are designed such that they provide relatively independent signals across grades. As long as these signals are associated with future labor market returns, than focusing on teacher value-added may be beneficial. Differentiating between these hypotheses is left for future research.

The other prominent result in Table 4 is the significant difference in the estimated standard deviation of teacher quality. Estimating the discount rate directly yields an estimate of the standard deviation in teacher quality equal to 0.198. If I assume that teacher inputs do not decay at all the standard deviation of teacher quality is 0.249, a bias of approximately 25%. This is similar in sign and magnitude to the bias detected in the simple monte carlo exercises discussed in Section 2.

In order to determine the overall effect of teachers in the production of achievement, the standard deviation of teacher quality must be compared relative to the standard deviation of student test scores. A one standard deviation increase in teacher quality results in 33% of a standard deviation increase in the growth of test scores, or approximately 20% of a standard deviation increase in test score levels. Not surprisingly, this last result is significantly larger than previous estimates. In much of the previous literature value-added models were estimated in levels and assumed that teacher inputs decayed fully from one year to the next. As the monte carlo exercises earlier in the paper illustrate, if the decay rate is not in fact zero, the estimated variance in teacher quality will be biased downward.

To put the impacts of persistence and variation in teacher quality in perspective, consider two students with mean ability equal to zero. The first student is assigned a teacher two standard deviations above the mean in grades three through five. The second student is assigned an average teacher in grades three and four, and a teacher two standard deviation above the mean in fifth grade. On average, student one will score approximately 10% of a standard deviation higher in fifth grade strictly as result of having better teachers in the past.

One final point regarding the results from these two basic models is that the teacher value-added estimates across the two models are very highly correlated, around 0.95. This suggests that in general the no decay model does a decent job in terms of ranking teachers. For example, consider a principal who wants to identify the worst teacher in a particular grade. The models give a different prediction only 18% of the time. Thus, while it is imperative to account for the persistence of inputs to gauge how important teachers are in the production process overall, the simpler model still provides a useful ranking of teachers.

4.2 Heterogeneity and Classroom Attributes

The results discussed thus far assume a constant geometric discount rate and a lack of other classroom attributes correlated with both teacher quality and the growth in test scores. Tables 5 and 6 relax these assumptions. Table 5 shows results from a model where I allow the discount rate to vary across third and fourth grade. The results indicate that in fact teacher inputs in fourth grade persist at a greater rate than teacher inputs in third grade. This could indicate that fourth grade is a sensitive period in the production of human capital, or it might simply reflect variability in the usefulness of each curriculum going forward. A standardized test across grades would help separate out these two theories. If it turns out that fourth grade is a sensitive period, this may have implications for how principals assign teachers to grades. However, I do not want to put too much emphasis on these results since there is only one future period with which to estimate the decay rate associated with fourth grade. Longer panels are better suited for identifying heterogeneity in the discount rate.

The first column of Table 6 shows results when I again assume a constant geometric discount rate for past teacher inputs, but allow other classroom attributes to have both a contemporaneous and persistent effect. I allow the discount rate on other classroom attributes to differ from the discount rate on teacher inputs. The results indicate that both teacher input and other classroom attributes decay at very similar rates, about 0.25. The coefficients on the classroom attributes largely coincide with previous findings. Class size has a small, but significant negative effect on the growth in student test scores. Students who transfer schools as part of a typical path through the school system actually increase their gains. This likely reflects differences in school quality across different schools and districts. The proportion of non-white peers and male peers both have negative but insignificant effects.

The final row of Table 6 shows that the standard deviation in teacher quality is slightly higher in this model. If higher quality teachers are assigned larger classes or more troubled students we would have expected this number to increase. The fact that it does not is not all that surprising since teacher satisfaction depends to a large extent on the classroom environment. If the best teachers are constantly assigned the biggest classes with the worst kids, the likelihood of losing these teachers will probably be much higher.

4.3 Varying Teacher Quality

Finally, I want to incorporate time-varying teacher attributes into the model to allow for overall quality to evolve over time. I separate out this version of the model from the previous versions since I can only estimate this model on the much smaller match sample. The only time-varying characteristic included is teacher experience. Over the short time frame considered there is very little variation in teacher education (masters versus bachelors degree) or certification. In practice with a longer panel it would be possible to incorporate these components as well.

Following much of the literature, rather than include a linear teacher experience term, I include a set of dummy variables that indicate if a teacher is in one of four categories: no experience, one year of experience, two years of experience, or between three and five years of experience.[15] It is widely believed that after five years there is essentially no gain to experience and the production function is constructed to reflect that.

The results are contained in the second column of Table 6. In this specification I assume

[15] Clotfelter et al. (2007)

that the discount rate for all past inputs are identical given the similarity in the coefficient estimates in the previous section. The estimates of the discount rate and the coefficients on the other classroom attributes are quite similar between this model and the specification excluding experience. The teacher experience variables follow the expected pattern and are highly statistically significant. As teachers gain experience, their impact on student performance increases greatly over the first few years on the job. To help put the experience coefficients in perspective, an average teacher with zero years of experience is approximately equivalent to a teacher with more than five years of experience but two standard deviations down the teacher quality distribution.

A nice feature of the model is that it isolates teacher quality independent of actual teacher experience. Thus we can examine if teachers with more experience are negatively selected. Regressing the teacher value-added estimates on teacher experience, weighting by the variance of the individual teacher quality estimates, yields a coefficient equal to -0.001 that is statistically significant at a 1% level. In other words, a teacher with 20 years of experience is approximately 10% of a standard deviation lower in the quality distribution than a new hire. So while the high discount rate suggests that rewarding teachers based strictly on value-added may not be optimal, rewarding teachers based entirely on experience also appears to be a sub-optimal strategy.

5 Conclusion

In this paper I develop and estimate a model of student achievement that explicitly accounts for both the accumulation of teacher inputs over time and the non-random sorting of students to teachers. I find that teachers play a significant role in the contemporaneous production of student achievement, but that in fact teacher value-added decays at a very high rate. There is some evidence that compared to third grade, fourth grade is a sensitive period in the production of knowledge, though lengthier panels of test scores would be useful in teasing this relationship out. Teacher experience continues to play an important role in the production process. As teachers gain experience, they become much more productive. However, on average, more experienced teachers are lower in the quality distribution suggesting a negative selection effect.

As discussed in the text, there are a number of ways to interpret the results regarding teacher effects. The first interpretation, that teacher value-added is a poor measure for the long-term impact of teachers, suggests that evaluating teachers using this methodology may be misguided. The second interpretation, that the tests and curriculum are designed such that each test provides an independent signal may mean that evaluating teachers through value-added is fruitful, particularly since there appears to be significant variation in teacher quality. Differentiating between these hypothesis will be critical and is left for future work.

References

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

Arcidiacono, P., G. Foster, N. Goodpaster, and J. Kinsler, "Estimating Spillovers in the Classroom with Panel Data," 2006. Unpublished Manuscript.

Boardman, A., , and R. Murname, "Using panel data to improve estimates of the determinants of educational achievement," *Sociology of Education*, 1979, *52* (2), 113–121.

Clotfelter, C., H. Ladd, and J. Vigdor, "Teacher-Student Matching and the Assessment of Teacher Effectiveness," *Journal of Human Resources*, 2006.

_ , _ , and _ , "How and Why Do Teacher Credentials Matter For Student Achievement," 2007. NBER Woking Paper 12828.

Goldhaber, D. and D. Brewer, "Does Teacher Certification Matter? High School Teacher Certification Status and Student Achievement," *Educational Evaluation and Policy Analysis*, 2000.

Hanushek, Eric, John Kain, and Steven Rivkin, "Teachers, Schools, and Academic Achievement," *Econometrica*, 2005, *73* (2), 417–458.

Harris, D. and T. Sass, "Teacher Training, Teacher Quality and Student Achievement," 2006. Woking Paper.

_ and _ , "Value-Added Models and the Measurement of Teacher Quality," 2006. Woking Paper.

Jacob, Brian, Lars Lefgren, and David Sims, "The Persistence of Teacher Value-Added," 2007. Woking Paper.

Rivkin, Steven, "Cumulative Nature of Learning and Specification Bias in Education Research," 2006. Unpublished Manuscript.

Rockoff, Jonah, "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data," *American Economic Review*, 2004, pp. 247–252.

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

Todd, P. and K. Wolpin, "The Production of Cognitive Achievement in Children: Home, School, and Racial Test Score Gaps," 2006. Woking Paper.

Table 1: Bias in Simple Persistence Models

	Actual $(1-\delta)$	True Var(T)	Var$(\hat{T})$
Assume: $(1-\delta)=1$	0.25	0.063	0.108
	0.5	0.063	0.094
	0.75	0.062	0.078
Assume: $(1-\delta)=0$	0.25	0.062	0.0526
	0.5	0.062	0.0441
	0.75	0.062	0.037

Table 2: Small Sample Performance of Various Accumulation Models

	$(1-\delta_1)$	$(1-\delta_2)$			$Var(T)$	Adj. $Var(\hat{T})$
Actual Value	**0.25**	**0.15**				
Baseline	0.253				0.246	0.251
	(0.018)					
Grade Specific δ's	0.249	0.146			0.251	0.253
	(0.020)	(0.040)				
Non-Geometric δ's	0.243	0.135			0.25	0.252
	(0.039)	(0.071)				

	$(1-\delta_1)$	$(1-\delta_2)$	β_1	β_2	$Var(T)$	Adj. $Var(\hat{T})$
Actual Value	**0.25**	**0.1**	**-0.05**	**-0.2**		
Class X's	0.252	0.101	-0.050		0.251	0.253
	(0.019)	(0.023)	(0.001)			
Class and Teacher X's	0.249	0.098	-0.050	-0.200	0.251	0.253
	(0.008)	(0.029)	(0.001)	(0.007)		

Table 3: Student Data

	Population		Base Sample		Match Sample	
Number of Students	452,922		144,236		39,376	
Number of Schools	1,353		633		260	
Number of Teachers	29,433		14,331		3,928	
	Mean	SD	Mean	SD	Mean	SD
Student Characteristics						
Male	0.51		0.51		0.51	
Non-White	0.42		0.36		0.28	
Transfer	0.08		0.04		0.02	
Structural Transfer	0.05		0.02		0.02	
Class Size	22.5	3.55	23.01	3.23	23.38	3.02
Peer Male	0.51	0.09	0.51	0.08	0.51	0.08
Peer Non-White	0.42	0.29	0.36	0.27	0.28	0.22
Math Test Scores						
Grade 2	0.12	0.98	0.27	0.93	0.33	0.92
Grade 3	1.98	0.87	2.13	0.82	2.19	0.82
Grade 4	2.75	0.97	2.89	0.93	2.96	0.93
Grade 5	3.23	1.12	3.41	1.08	3.52	1.07
Teacher Statistics						
Avg. # of Students	40.5	27.7	30.2	21.3	30.1	17.8
Avg. # of Years Observed	1.87	1.22	1.92	1.22	2.63	1.26
Teacher Experience					14.1	9.6

Table 4: Baseline Model Results

	No Decay	Baseline
$(1-\delta)$	-	0.233*
		(0.018)
SD(Teacher Quality)	0.249	0.198
Total Teachers	14,331	14,331
Student Heterogeneity	Y	Y
Grade Fixed Effects	Y	Y
Sample Size	432,708	432,708

Table 5: Heterogeneity in δ

$(1-\delta_{\text{3rd Grade}})$	0.207*
	(0.017)
$(1-\delta_{\text{4th Grade}})$	0.368*
	(0.036)
SD(Teacher Quality)	0.202
Total Teachers	14,331
Student Heterogeneity	Y
Grade Fixed Effects	Y
Sample Size	432,708

Table 6: Include Classroom and Teacher Attributes

$(1-\delta)$ - Teacher Discount	0.240*	0.218*
	(0.018)	(0.035)
$(1-\gamma)$ - Classroom Discount	0.264*	
	(0.043)	
Transfer	-0.003	-0.003
	(0.005)	(0.016)
Structural Transfer	0.033*	0.062*
	(0.017)	(0.028)
Class Size	-0.005*	-0.006*
	(0.001)	(0.001)
% Peer Non-White	-0.003	0.003
	(0.013)	(0.023)
% Peer Male	-0.006	-0.024
	(0.016)	(0.031)
No Experience		-0.374*
		(0.022)
1 yr. of Experience		-0.219*
		(0.019)
2 yrs. of Experience		-0.127*
		(0.017)
3-5 yrs. of Experience		-0.053*
		(0.012)
SD(Teacher Quality)	0.200	0.184
Total Teachers	14,331	3,928
Student Heterogeneity	Y	Y
Grade Fixed Effects	Y	Y
Sample Size	432,708	118,128