Beyond Levels and Growth

Estimating Teacher Value-Added and its Persistence

Josh Kinsler

A B S T R A C T

The levels and growth achievement functions make extreme and diametrically opposed assumptions about the rate at which teacher inputs persist. I first show that if these assumptions are incorrect, teacher value-added estimates can be significantly biased. I then develop a tractable, cumulative model of student achievement that allows for the joint estimation of unobserved teacher quality and its persistence. The model can accommodate varying persistence rates, student heterogeneity, and time-varying teacher attributes. I implement the proposed methodology using schooling data from North Carolina, and find that only a third of the contemporaneous teacher effect survives into the next grade.

I. Introduction

Teacher quality is widely believed to be the most important school-level input into the production of student achievement. However, quantifying the amount of variation in student test scores that can be attributed to differing teacher assignments is difficult since teacher ability is largely unobservable.[1] To overcome this hurdle, researchers often treat teacher quality as an unobserved parameter to be estimated directly using student test-score variation across classrooms and over time.[2] This approach, widely known as value-added modeling, has two primary benefits. First, value-added models provide an objective, teacher-specific measure of

1. Observable teacher characteristics, such as education, experience, and licensure, have been shown to have limited predictive power for student outcomes. See Hanushek and Rivkin (2006) for a review of this literature.

2. Sanders, Saxton, and Horn (1997) were the first to develop and employ a value-added framework.

Josh Kinsler is an assistant professor of economics at the University of Rochester. *The author would like to thank seminar participants at NYU, the University of Western Ontario, and participants at the 34th Annual AEFP Conference. The author also wishes to thank Peter Arcidiacono, Ronni Pavan, and Nese Yildiz for helpful comments. Researchers may acquire the data used in this article from the North Carolina Education Research Data Center http://childandfamilypolicy.duke.edu/project_detail.php?id = 35. For questions regarding the data, contact Josh Kinsler, joshua.kinsler@rochester.edu.*

[Submitted November 2010; accepted July 2011]

ISSN 022-166X E-ISSN 1548-8004

effectiveness that many favor over subjective measures such as principal or peer evaluations. Second, value-added models allow researchers to trace out the entire distribution of teacher quality, which is useful for determining how much of the gaps in student outcomes can be attributed to variable school inputs. Although the benefits of the value-added approach are clear, considerable debate remains about whether these models can consistently estimate teacher effectiveness.

The controversy surrounding the use of value-added models is in part related to the fact that there is no benchmark methodology. Value-added modeling is a broad term that encompasses a variety of approaches that share one common feature, measuring the effectiveness of individual teachers using student test scores.[3] However, the various approaches within the broader value-added framework often make conflicting assumptions regarding the underlying achievement production function. In particular, assumptions about how teacher inputs persist over time are often diametrically opposed. As an example, three of the most widely cited papers that estimate the distribution of teacher quality make drastically different assumptions regarding the persistence of teacher inputs. Rockoff (2004) assumes that teacher inputs do not persist at all, identifying teacher effectiveness using variation in the level of student test scores. Hanushek, Kain, and Rivkin (2005) make the exact opposite assumption, perfect persistence from one year to the next, when they use variation in test-score growth to estimate contemporaneous teacher effectiveness. Finally, Aaronson, Barrow, and Sander (2007) take a middle ground and assume that teacher inputs, along with all other inputs including student ability, persist at a constant geometric rate.

Despite the significant differences in approach, the basic findings across all three aforementioned studies are quite similar. A one standard deviation increase in teacher quality yields approximately 10 percent of a standard deviation increase in math test scores and slightly smaller effects in reading. The obvious question given the similarity in results across the various specifications is whether the assumptions about teacher persistence actually matter. Using theoretical and empirical examples I show that incorrect persistence assumptions can lead to significant biases in estimates of individual teacher value-added and dispersion in teacher value-added. The magnitude of the biases depend on how students and teachers are sorted across schools and classrooms and how poor the initial persistence assumptions are. In particular, the bias in estimates of teacher value-added tend be large when the true teacher value-added experienced by each student is highly correlated across grades, either positively or negatively. Additionally, the growth model will tend to perform well when the true persistence rate is close to one, while the levels model will tend to perform well when the true persistence rate is close to zero.

Uncertainty about the actual rate at which teacher inputs persist has led to a number of recent papers that estimate the depreciation rate directly. In a lag score model, Jacob, Lefgren, and Sims (2010) and Kane and Staiger (2008) show that using either an indicator or the actual measure of the lagged teacher quality as an instrument for the lagged test score yields an estimate of teacher persistence. Across the two papers, estimates of teacher persistence range between 0.15 and 0.5. How-

3. Throughout the paper I will use the terms teacher value-added, teacher effectiveness, and teacher quality interchangeably.

ever, neither paper estimates the decay rate jointly with teacher quality, which can lead to potential biases in both parameters. Lockwood et al. (2007) jointly estimate teacher quality and the rate of persistence using Bayesian MCMC techniques. Estimated decay rates are quite small; however, the model does not account for any observed or unobserved variation in student quality. Finally, Rothstein (2010) estimates a cumulative model that allows teachers to have separate contemporaneous and future effects.[4] The rate of persistence, estimated between 0.3 and 0.5, can be recovered in a second step by comparing the average relationship between the contemporaneous and future effect for each teacher. However, computational challenges severely limit both the estimation sample and production technology employed.

In this paper, I add to the above literature by developing and estimating a simple, but comprehensive, cumulative production technology that jointly estimates teacher quality and persistence. The proposed achievement model is flexible along many dimensions. Individual teacher quality is treated as a time-invariant unobserved parameter to be estimated; however, teacher effectiveness can vary over time as teachers accumulate experience and additional schooling. The rate at which teacher inputs persist can be geometric, or can vary across grades or over time in a more flexible fashion. The model can be estimated in either achievement levels or growth and can accommodate unobserved student heterogeneity in ability in either context. Contemporaneous and lagged observable student and classroom attributes also can be incorporated.

Despite the potentially large number of teacher and student parameters, the model can be computed in a timely manner. Rather than estimate all the parameters in one step, I take an iterative approach that toggles between estimating teacher persistence, teacher heterogeneity, student heterogeneity, and any remaining parameters.[5] Each iteration is extremely fast since I use the first-order conditions generated from minimizing the sum of the squared prediction errors to estimate the teacher and student heterogeneity directly. The iterative procedure continues until the parameters converge, at which point the minimum of the least squares problem has been achieved. With a sample of over 600,000 students and 30,000 teachers I can estimate the baseline model in less than 15 minutes.

Using student data from North Carolina's elementary schools I implement the proposed cumulative model and find that teacher value-added decays quickly, at rates approximately equal to 0.35 for both math and reading. A one standard deviation increase in teacher quality is equivalent to 24 percent of a standard deviation in math test scores and 14 percent of a standard deviation in reading test score. These results are consistent with previous evidence regarding the magnitude of teacher value-added and the persistence of teacher inputs. Using the levels or growth frameworks instead results in biases in the variance of teacher value-added on the order of 4 percent and 7 percent respectively for math test-score outcomes, and smaller biases in reading outcomes. However, the somewhat small biases in the variance of overall teacher value-added masks larger within-grade biases that are on the order

4. Carrell and West (2010) take a similar approach when measuring the short- and long-term effectiveness of professors at the college level. In contrast to the findings from Rothstein (2010) and Jacob, Lefgren, and Sims (2010) at the elementary school levels, they find that the short- and long-term effects are actually negatively correlated.

5. This approach is similar in spirit to the methodology outlined in Arcidiacono et al. (Forthcoming).

of 15 percent. The individual estimates of teacher quality from the cumulative model and the levels model are very highly correlated, while the growth model estimates individual teacher value-added less accurately. In general, the similarities in teacher value-added across the cumulative, levels, and growth models in the North Carolina sample reflect the fact that for the average student, teacher value-added is only marginally correlated across grades.

The remainder of the paper is as follows. The pitfalls of the levels and growth frameworks are illustrated in Section II. Section III outlines a cumulative production function that allows for flexible persistence patterns, discusses identification of the key parameters, and provides an estimation methodology. Section IV introduces the North Carolina student data used to estimate the cumulative production function. Section V contains analysis of the model results and Section VI concludes.

II. Persistence Assumptions and Estimates of Teacher Quality

The most common value-added models of teacher quality assume that teacher effects either persist forever or not at all. The motivation for making these extreme assumptions is typically model tractability. Consider the following two achievement equations:

$$(1) \quad A_{ijg} = \alpha_i + \sum_{j \in J_g} I_{ijg} T_{jg} + \varepsilon_{ijg}$$

$$(2) \quad A_{ijg} - A_{ij'g'} = \sum_{j \in J_g} I_{ijg} T_{jg} + e_{ijg}$$

where A_{ijg} is the achievement outcome for student i matched with teacher j in grade g. J_g is the set of all grade g teachers and I_{ijg} is an indicator function that takes on the value of one if student i is matched with teacher j in grade g. T_{jg} is the value-added of teacher j in grade g and α_i is the grade-invariant ability level of student i.

Both Equations 1 and 2 allow for unobserved student ability to affect the level of individual test scores. Equation 1 is a simplified version of the levels specification employed in Rockoff (2004), which implicitly assumes that grade g's outcome is entirely unaffected by teachers in previous grades. Equation 2 is a growth equation that can be generated by first differencing two levels scores. However, in order for only the contemporaneous teacher to enter in the equation, past teacher inputs must persist perfectly. The key benefit to either of the extreme assumptions about teacher persistence is that only one set of teacher effects enters into Equations (1) and (2), making estimation relatively straightforward. However, as I show below, these extreme assumptions can have significant consequences for both $\hat{T}_{jg}$ and $\hat{\sigma}^2_T$.

I start with a simple theoretical example that illustrates the types of biases that arise in the levels and growth models when in fact teacher inputs persist at a particular rate δ. I follow this up with some simple Monte Carlo exercises to demonstrate the size of the bias under various teacher and student sorting scenarios. In the following sections, I do not separately investigate the lag-score model applied by Aaronson, Barrow, and Sander (2007) since I view it as a special case of either the

growth or levels framework. However, because the lag-score model is typically implemented without controls for unobserved student heterogeneity, the coefficient on the lag score is typically quite high, on the order of 0.8. This implies that past teacher inputs also decay at a rate equal to 0.8. Thus, I suspect that the pattern of biases in the lag-score framework are best approximated by the biases in the growth model.

A. Theoretical Example

Assume that we observe three test-score outcomes for each of three different students, denoted A_{ig} where i indexes students and g indexes grade or time. The three students are all members of the same school. The first test-score observation for each student is not associated with any teacher and can be considered an unbiased measure of a student's unobserved ability. This assumption is useful since it ensures that all the student and teacher parameters will be separately identified and is also consistent with the data to be used in the empirical analysis. In the remaining two grades, students are assigned to one of two teachers that are unique to each grade, where T_{jg} is the jth teacher in grade g. Thus, there are four teachers in the school: T_{12}, T_{22}, T_{13}, and T_{23}. As an example, the set of outcomes for student 1 takes the following form:

$$A_{11} = \alpha_1 \tag{3}$$

$$A_{12} = \alpha_1 + T_{12} \tag{4}$$

$$A_{13} = \alpha_1 + T_{13} + \delta T_{12} \tag{5}$$

In order for δ to be identified, students must switch classmates in Grade 3, generating variation in T_{j2} for one of the Grade 3 teachers. I assume that students 2 and 3 are assigned teacher T_{22} in Grade 2 and students 1 and 2 are assigned teacher T_{13} in Grade 3. The singleton student classes are matched with teachers T_{12} and T_{23}. All of the unobserved parameters, α_i, δ, and T_{jg} are identified in this simple system. Since I have written the achievement outcomes without any measurement error, it is possible to pin down exactly each of the unobserved parameters. The question is what happens to the estimates of $\hat{T}_{jg}$ and $\hat{\sigma}^2_{jg}$ when we assume that δ equals either 0 or 1.

To derive the least squares solutions for $\hat{T}_{jg}$ when δ is assumed to equal 0, I simply differentiate the squared deviations and solve for the parameters as a function of A_{ig}, the only observables in the model. I then substitute back in the true data generating process for A_{ig} to illustrate how the estimates differ from the true underlying parameters. In the levels case, $\delta = 0$, the least squares estimates of the four teacher effects is given by

$$\hat{T}^L_{12} = T_{12} - \frac{4\delta}{15}(T_{12} - T_{22})$$

$$\hat{T}^L_{22} = T_{22} + \frac{2\delta}{15}(T_{12} - T_{22})$$

$$\hat{T}^L_{13} = T_{13} + \frac{\delta}{15}(7T_{12} + 8T_{22})$$

$$\hat{T}^{L}_{23} = T_{23} + \frac{\delta}{15}(T_{12} + 14T_{22})$$

The first thing to notice is that unless δ equals zero all of the estimates of teacher quality will be biased, even the Grade 2 teacher effects. The bias in the Grade 2 teacher effects arises because the unobserved student abilities adjust to account for the unexplained variance in the third grade outcome. This in turn leads to bias in all the teacher effect estimates. The teacher effects in second grade are always biased toward the average effect, while the bias in the third grade teacher effects depend on whether or not T_{12} and T_{22} are greater than zero. If the second grade teachers are "good," then the estimated effects of the Grade 3 teachers will be biased upward since they receive some credit for the lingering effects of the previous grade's teacher. The opposite result is true if the second grade teachers are "bad." Thus, teachers following excellent teachers will be unjustly rewarded while teachers following a string of poor teachers are unjustly penalized in a model that incorrectly assumes no persistence.

Not only are the individual estimates of teacher quality biased, but the overall distribution of teacher quality is also affected. Because the Grade 2 teacher effects are always biased toward the average effect, $\hat{\sigma}^2_{j2}$ will be biased toward zero. The bias in the estimated dispersion of the Grade 3 teacher effects is given by

$$(6) \quad \hat{\sigma}^2_{j3} - \sigma^2_{j3} = \frac{4\delta^2}{50}(T_{12} - T_{22})^2 + 20\delta(T_{13} - T_{23})(T_{12} - T_{22})$$

The first term on the right hand side of the above expression is positive for any rate of persistence. However, the second term can be either positive or negative depending on the correlation in teacher quality across time periods. In this simple example the cross-grade correlations in teacher quality are weighted equally. However, with more students and teachers, the weight given to each cross-grade correlation will depend on how many students are associated with each teacher pair. Thus, depending on how students are tracked through classes, the bias in $\hat{\sigma}^2_{j3}$ could be either positive or negative. If the bias in the the variance of the third grade teacher effects is large enough, it could swamp the bias in the estimate of teacher dispersion in second grade, leading to an inconclusive overall bias in the variance of teacher quality.

Rather than assume $\delta = 0$, we could instead assume that $\delta = 1$ and estimate a simple growth score model with two observations per student. In this case the biases work quite differently. Because the student fixed effects are differenced out of the model, there is nothing to cause bias in the Grade 2 teacher effects. As a result, $\hat{T}^{G}_{j2} = T_{j2}$.[6] However, the estimated Grade 3 teacher effects will be biased as a result of the incorrect assumptions about the persistence of the Grade 2 teachers. In particular,

$$\hat{T}^{G}_{13} = T_{13} + \frac{1}{2}(\delta - 1)(T_{12} + T_{22})$$

6. If we allowed for unobserved student heterogeneity in the growth of student test scores this result would no longer hold.

$$\hat{T}_{23}^{G} = T_{23} + (\delta - 1)T_{22}$$

Clearly if $\delta = 1$ neither teacher effect is biased. Otherwise, both Grade 3 teacher effects will be biased in a direction that again depends on the quality of the Grade 2 teacher. In contrast to the levels model, following an excellent Grade 2 teacher will bias downward the Grade 3 teacher effect since excess credit is given to the previous teacher. Thus, the teachers that get unjustly rewarded in the levels model are the same teachers who get unjustly punished in the growth framework.

The bias in the estimates of the Grade 3 teacher effects bleeds into the estimate of the overall dispersion of teacher quality. The bias in the estimated variance of Grade 3 teacher quality is given by

$$(7) \qquad \hat{\sigma}_{j3}^2 - \sigma_{j3}^2 = \frac{1}{8}((\delta - 1)^2(T_{12} - T_{22})^2 + 4(\delta - 1)(T_{13} - T_{23})(T_{12} - T_{22}))$$

The first term inside the parentheses is always positive while the sign of the second term will depend on the cross-grade covariances in teacher quality. Again, in a model with more students and teachers, the weight given to the cross-grade teacher covariance terms will depend on the number of students associated with each teacher pair. If teacher ability is correlated across grades at the student level, the dispersion in teacher quality in the growth framework will be understated since $(\delta - 1)$ is negative. This is the opposite of the bias in the levels framework.

The simple example above highlights a number of important issues regarding the standard levels and growth models employed to evaluate teachers. First, when teacher effects persist, the levels model yields biased estimates of all teacher effects, including the initial teacher. Second, the bias in the estimated dispersion in teacher quality varies significantly across grades, particularly in the levels framework. Finally, the bias in the individual estimates of teacher quality and overall dispersion in teacher quality depend critically on how students and teachers are tracked across grades. The next section illustrates this final point by examining empirically the magnitude of the biases under various sorting scenarios.

B. Empirical Exercise

The purpose of this section is to illustrate the magnitude of the biases in the dispersion of teacher quality under various teacher and student sorting scenarios. I expand the simple model from the previous section and generate a data set that is representative of the type of schooling data available to researchers. I assume that there are 25 schools in the sample and 5,000 students. For each student, a baseline test-score measure is available, followed by three classroom test-score observations associated with a particular grade, say Grades 3, 4, and 5. Within each school, there are four teachers per grade, and each teacher is observed with 50 students. I continue to assume that the achievement tests are perfect measures of student knowledge, primarily to illustrate that any biases that emerge are strictly a result of model misspecification.

Consistent with the previous section, I assume that the true underlying achievement production function is cumulative in teacher inputs. Past teacher inputs are assumed to persist at a rate equal to 0.5. Moving the persistence rate closer to 0 or 1 will improve either the levels or growth model at the expense of the other. I

assume that student ability is distributed N(0,1) and that teacher quality is distributed N(0,0.0625).[7] The distribution of teacher quality is identical across the three grades.

The final and most important decision when generating the data is how to distribute students and teachers to schools and classrooms. As the previous section highlights, the bias in the dispersion of teacher quality will depend to a large extent on the cross-grade correlations in teacher quality at the student level. There are two methods for generating cross-grade correlations in teacher quality: sorting of students to teachers within schools, and sorting of teachers across schools. Within schools, if individual students are consistently matched with either relatively high- or low-ability teachers, then teacher quality will be positively correlated across grades. A negative correlation would result if principals, in an attempt at fairness, assign a relatively high-ability teacher one year followed by a relatively low-ability teacher the next. Even if there exists no sorting within schools, teacher ability will correlated across grades if teachers sort across schools based on ability. Regardless of the source of the cross-grade correlation in teacher quality, the previous section suggests that the estimated dispersion in teacher quality will be significantly impacted.[8]

Table 1 illustrates how the estimated dispersion in teacher quality from both the levels and growth models are affected by varying amounts of cross-grade correlation in teacher ability.[9] Within each model type, levels, or growth, the first row contains the estimates of teacher dispersion when there is no cross-grade correlation in teacher quality, followed by three levels of positive and negative sorting.[10] The results are averages over 500 simulations where the underlying population of students and teachers is held fixed, and only the sorting of students to teachers within schools changes with each iteration.

The pattern of results is quite consistent with the predictions from the simple theoretical exercise. The bias in the estimated variance of teacher quality varies significantly across grades for all types of sorting, particularly in the levels framework. In the growth model, the variance of third grade teacher quality is always perfectly identified, regardless of how students and teachers are sorted. The directions of the bias in the levels and growth models is as expected. With no cross grade correlation in teacher quality, the overall variance in teacher value-added is biased downward in the levels framework and upward in the growth framework. These biases are generated by the fact that past teachers are an omitted variable in the contemporaneous outcomes. With negative sorting these patterns are exaggerated, and with significant positive sorting these biases are flipped. With extreme levels of positive (negative) sorting, the estimated variance of teacher value-added is biased upward in the levels (growth) model by 72 percent (43 percent).

7. The dispersion in student ability and teacher quality is similar to those estimated using data from North Carolina.

8. If an unbiased measure of student ability were unavailable, only within-school variation in teacher quality would be identified. In this case the source of the cross-grade correlation in teacher quality would matter since if it was generated all by teacher sorting across schools, the within-school estimates of dispersion would be unaffected.

9. I chose the cross-grade correlations in an effort to illustrate the performance of the levels and growth models across a spectrum of potential realizations. For the North Carolina data, it turns out that the cross-grade correlation in teacher quality is approximately 0.11.

10. Note that I relied solely on within school sorting to generate these correlations. The results would be unchanged if I also used teacher sorting across schools to generate similar levels of cross-grade correlation in teacher quality.

Table 1
Estimating Variance in Teacher Ability using Misspecified Levels and Growth Models

True Persistence of Teacher Ability: $\delta = 0.5$					
Model Type	$\rho(T_{jg},T_{j,g+1})$	σ^2_T	$\sigma^2_{T_{j3}}$	$\sigma^2_{T_{j4}}$	$\sigma^2_{T_{j5}}$
		0.063	**0.062**	**0.063**	**0.063**
Levels—($\delta = 0$)	**0.00**	0.054	0.042	0.052	0.068
	0.12	0.057	0.040	0.057	0.075
	0.25	0.063	0.040	0.064	0.085
	0.58	0.107	0.062	0.115	0.146
	−0.12	0.052	0.045	0.049	0.062
	−0.25	0.050	0.047	0.045	0.057
	−0.58	0.053	0.062	0.042	0.055
Growth—($\delta = 1$)	**0.00**	0.066	0.062	0.067	0.068
	0.12	0.060	0.062	0.059	0.058
	0.25	0.055	0.062	0.053	0.049
	0.58	0.048	0.062	0.042	0.038
	−0.12	0.070	0.062	0.075	0.072
	−0.25	0.075	0.062	0.085	0.077
	−0.58	0.090	0.062	0.115	0.092

Note: Results are averages over 500 simulations. Generated sample contains 25 schools, 3 grades per school, and 4 teachers per grade. Each teacher is observed with 50 students. Student and teacher ability are unobserved. Grades are generated according to a cumulative achievement equation where teacher inputs persist at a constant geometric rate equal to 0.5. Students are observed four times, first without any associated teacher, and then once in each grade. There is no additional measurement error in the model so that any biases stem entirely from model misspecification. Teacher and student populations are held fixed across the simulations, with only the within school sorting of students to teachers changing. T_{jg} is the ability of teacher j in grade g. $\rho(T_{jg},T_{j,g+1})$ is the correlation in teacher quality across grades. σ^2_T is the true variance of teacher ability across all grades, while $\sigma^2_{T_{j3}}$ is the true variance of ability for third grade teachers only. The levels model implicitly assumes a persistence rate of 0 while the growth model implicitly assumes a persistence rate of 1. Bold-faced numbers reflect true underlying distributions.

Not only are the estimates of the dispersion in teacher quality affected by the strong persistence assumptions inherent in the growth and levels framework, the individual estimates of teacher quality are also affected. For each of the simulations, I also calculated the proportion of schools that were able to successfully identify the best teacher within each grade. Under random assignment, both models actually perform well, identifying the best teacher in all grades more than 95 percent of the time. When the cross-grade correlation in teacher quality is equal to 0.25, the levels model continues to perform quite well, however the growth model identifies the best fourth and fifth grade only 83 percent and 74 percent of the time respectively. When the cross-grade correlation in teacher quality is equal to −0.25 performance is flipped, with the growth model almost always identifying the best teacher in each

grade, while the levels model only identifies the best teacher in fourth and fifth grade 87 percent of the time.

The intuition for the difference in the ability of the levels and growth models to identify the best teacher in each grade is the following. With positive sorting, the best (worst) fourth grade teacher tends to follow the best (worst) third grade teacher. Because the levels model assumes that the third grade teacher has no effect in fourth grade, the impact of the best (worst) fourth grade teacher will be biased upward (downward). This tends to increase the estimated dispersion in teacher quality, while maintaining the relative rank of each teacher. As a result, the correlation between the estimated teacher value-added and the truth is high. In the growth framework, if teacher quality is positively correlated across grades, the effect of the best (worst) fourth grade teacher is biased downward (upward). The best fourth grade teacher follows the best third grade teacher, who continues to get significant credit for fourth grade performance since it is assumed that teacher effects persist forever. The estimated dispersion in teacher quality will be reduced and the ranking of teachers becomes significantly more jumbled. The stronger the positive correlation and the closer δ is to zero the more jumbled the ranking would become. If the true correlation in teacher value-added across grades were negative, the opposite would hold. The estimated teacher value-added from the growth (levels) framework would be highly (lowly) correlated with the truth.

Overall the results from this simple empirical exercise illustrate that both the levels and growth models are likely to be biased, though the direction and the severity will depend on how teachers and students are sorted across schools and classrooms.[11] The concern is that the extent of sorting on unobserved teacher and student ability is unknown a priori, thus leaving researchers with little guidance about which model will work best in particular situations. Rather than make strong and potentially incorrect assumptions about the persistence of teacher quality, I develop a straightforward method of estimating both teacher value-added and the rate at which teacher inputs persist.

III. Estimating Persistence

In the following sections, I lay out a methodology for estimating a cumulative production function for student achievement that yields estimates of teacher quality and the rate at which teacher inputs persist. I start with a baseline model that contains only unobserved student ability, teacher effects, and the persis-

11. It is important to point out that the biases illustrated in Table 1 stem from a particular underlying production technology and assumption about the types of data available. If, for example, an initial baseline test score is unavailable, only variation in teacher quality within schools is identified. Similar identification restrictions occur if heterogeneity in the growth of student test scores is included in the model. In these cases, the biases stemming from the sorting of teachers across schools would not be present. However, the population dispersion in teacher quality would be significantly understated if the within-school variation in teacher quality is used as a proxy.

tence parameter.[12] Within this simple framework I discuss identification, estimation, consistency, and important assumptions regarding the assignment of students to teachers. I then extend the basic model to allow for heterogenous decay rates and time-varying teacher quality. Finally, I provide some Monte Carlo evidence that illustrates the accuracy of the proposed estimation methodology.

A. Baseline Model

The biases in the teacher-specific measures of effectiveness and the overall dispersion in teacher quality stemming from incorrect assumptions about input persistence can be avoided by instead jointly estimating teacher quality and the persistence of teacher inputs. Expanding on the simple cumulative specification outlined in Equations 3–5, assume that the true achievement production function is given by

$$A_{ijg} = \alpha_i + \sum_{j \in J_g} I_{ijg} T_{jg} + \sum_{g'=1}^{g-1} \delta^{g-g'} \Big(\sum_{j \in J_{g'}} I_{ijg'} T_{jg'} \Big) + \varepsilon_{ijg} \tag{8}$$

where I_{ijg} is an indicator function that equals one when student i is assigned teacher j in grade g, and J_g is the set of all grade g teachers. ε_{ijg} represents measurement error in the test's ability to reveal a student's true underlying knowledge, and is uncorrelated with unobserved student ability and teacher assignments in all periods. I return to the issue of exogeneity at the conclusion of this section.

The first summation in achievement equation is the effect of the contemporaneous teacher, while the second summation accounts for the discounted cumulative impact of all the teachers in grades $g' < g$. For simplicity, I assume here that the teacher inputs persist at a constant geometric rate δ, though I will relax this assumption later. The lingering impact of past teachers is assumed to be proportional to their contemporaneous impact. In practice, I could allow teachers to have separate contemporaneous and long-term effects rather than estimate a mean persistence rate, similar to Rothstein (2010). However, I choose to measure teacher effectiveness with one parameter that captures a mixture of both effects.

Prior to discussing estimation, it is useful to briefly describe the variation in the data that identifies δ. As long as students are not perfectly tracked from one grade to the next, conditional on the current teacher assignment there will be variation in the lagged teacher. In practice, perfect tracking is rarely employed, and thus identification is obtained.

Estimation, on the other hand, is complicated by both the multiple sets of high-dimensional teacher fixed effects and the inherent nonlinearity stemming from the interaction between the persistence parameter and the teacher effects. To deal with these issues, I pursue an iterative estimation strategy similar in spirit to the one outlined in Arcidiacono et al. (Forthcoming). Define the least squares estimation problem as,

12. Throughout the paper, I exclude school effects from the achievement model. School effects can easily be incorporated, however, interpretation is much simpler without them since no additional normalizations are necessary. By attributing all cross-school variation in test scores to teachers I likely overstate the true variation in teacher quality. However, attributing all the cross-school variation in test scores to the schools themselves would likely understate the variance of teacher quality. Estimates including school fixed effects indicate that the persistence parameter is largely unaffected by their inclusion.

$$(9)\quad \min_{\boldsymbol{\alpha},\mathbf{T},\delta} \sum_{i=1}^{N} \sum_{g=\underline{g}}^{\bar{g}} (A_{ijg} - \alpha_i - \sum_{j \in J_g} I_{ijg} T_{jg} - \sum_{g'=1}^{g-1} \delta^{g-g'} (\sum_{j \in J_{g'}} I_{ijg'} T_{jg'}))^2$$

where $\boldsymbol{\alpha}$, $\mathbf{T}$, and δ are the parameters of interest, and $\underline{g}$ and $\bar{g}$ are the minimum and maximum grades for which test score are available. The iterative estimation method toggles between estimating (or updating) each parameter vector taking the other sets of parameters as given. At each step in the process the sum of squared errors is decreased, eventually leading to the least squares solution.[13]

In practice, estimation proceeds according to the steps listed below. The algorithm begins with an initial guess for $\boldsymbol{\alpha}$ and $\mathbf{T}$ and then iterates on three steps, with the qth iteration given by:[14]

- Step 1: Conditional on $\boldsymbol{\alpha}^{q-1}$ and $\mathbf{T}^{q-1}$, estimate δ^q by nonlinear least squares.
- Step 2: Conditional on δ^q and $\mathbf{T}^{q-1}$, update $\boldsymbol{\alpha}^q$.
- Step 3: Conditional on $\boldsymbol{\alpha}^q$ and δ^q, update $\mathbf{T}^q$.

The first step is rather self-explanatory, however the second and third steps require some explanation since it is not clear what updating means. To avoid having to "estimate" all of the fixed effects, I use the solutions to the first-order conditions with respect to $\boldsymbol{\alpha}$ and $\mathbf{T}$ to update these parameters.

The derivative of the least squares problem with respect α_i for all i is given by the following

$$(10)\quad \sum_{g=\underline{g}}^{\bar{g}} (A_{ijg} - \alpha_i - \sum_{j \in J_g} I_{ijg} T_{jg} - \sum_{g'=1}^{g-1} \delta^{g-g'} (\sum_{j \in J_{g'}} I_{ijg'} T_{jg'}))$$

Setting this equal to zero and solving for α_i yields

$$(11)\quad \alpha_i = \frac{\sum_{g=\underline{g}}^{\bar{g}} (A_{ijg} - \sum_{j \in J_g} I_{ijg} T_{jg} - \sum_{g'=1}^{g-1} \delta^{g-g'} (\sum_{j \in J_{g'}} I_{ijg'} T_{jg'}))}{\bar{g} - \underline{g}}$$

which is essentially the average of the test-score residuals purged of individual ability. To update the full vector of abilities in Step 2, simply plug the qth estimate of δ and the $q-1$ step estimate of $\mathbf{T}$ into the N ability updating equations.

The strategy for updating the estimates of teacher quality in Step 3 is essentially identical to the one outlined for updating α. The key difference is that the first-order condition for T_{jg} is significantly more complicated. Teachers in grades $g < \bar{g}$ will

13. Matlab code for the baseline accumulation model and a simple data generating program can be downloaded at http://www.econ.rochester.edu/Faculty/Kinsler.html

14. My initial guess for α is the average student test score across all periods, and my initial guess for T is the average test score residual for each teacher after controlling for student ability. When guessing T, I use only the contemporaneous student outcomes. Note that as long as the initial guesses aren't all skewed in one direction (either too large or too small) the results are not sensitive to the starting values.

have not only a contemporaneous effect on achievement outcomes, but also an effect on student outcomes going forward. Thus, test-score variation in future grades aids in identifying the quality of each teacher. The first-order condition of the least squares problem in Equation 9 with respect to T_{jg}, for $\underline{g} < g < \bar{g}$, is

$$(12)\quad \sum_{i \in N_{ijg}} \Big[(A_{ijg} - \alpha_i - T_{jg} - \sum_{g'=1}^{g-1} I_{ijg'} T_{jg'}) + \sum_{g'=g+1}^{\bar{g}} \delta^{g'-g}(A_{ijg'} - \alpha_i - \sum_{j \in J_{g'}} I_{ijg'} T_{jg'} - \sum_{g''=1}^{g'-1} \delta^{g'-g''} I_{ijg''} T_{jg''})\Big]$$

where N_{ijg} are the set of all students who are assigned teacher j in grade g. The first term inside the summation over i comes from the contemporaneous effect of T_{jg}. The second term inside the summation accounts for the effect of T_{jg} in all subsequent grades. Notice that T_{jg} appears in the final summation of Equation 12 since at some point g'' will equal g.

Setting the first-order condition with respect to T_{jg} equal to zero and solving for T_{jg} yields

$$(13)\quad T_{jg} = \frac{\sum_{i \in N_{ijg}} [A_{ijg} - \alpha_i - \sum_{g'=1}^{g-1} I_{ijg'} T_{jg'}]}{\sum_{i \in N_{ijg}} (1 + \sum_{g'=g+1}^{\bar{g}} \delta^{2(g'-g)})} + \frac{\sum_{i \in N_{ijg}} [\sum_{g'=g+1}^{\bar{g}} \delta^{g'-g}(A_{ijg'} - \alpha_i - \sum_{j \in J_{g'}} I_{ijg'} T_{jg'} - \sum_{g''=1, g'' \neq g}^{g'-1} \delta^{g'-g''} I_{ijg''} T_{jg''})]}{\sum_{i \in N_{ijg}} (1 + \sum_{g'=g+1}^{\bar{g}} \delta^{2(g'-g)})}$$

where I split the term into two pieces for ease of presentation even though they share a common denominator. Again, the first term uses information from grade g to identify T_{jg}, while the second term uses information from grades $g' > g$. Notice that the equation for T_{jg} includes $T_{jg'}$ for $g' \neq g$. Thus, updating T_{jg} in Step 3 requires substituting in the qth iteration estimates of $\boldsymbol{\alpha}$ and δ, as well as the $q-1$ step estimates of $T_{jg'}$ into Equation 13.[15]

Iterating on Steps 1–3 until the parameters converge will yield the $\hat{\alpha}$, $\hat{\mathbf{T}}$, and $\hat{\delta}$ that solve the least squares problem outlined in Equation 9. At this point it is useful to step back and consider whether these are consistent estimators of the underlying parameters of interest. Because each student is observed at most $\bar{g} - \underline{g}$ times, the $\hat{\alpha}$ will be unbiased, but inconsistent. This is a standard result in panel data models with large N fixed T, where N refers to the number of students and T refers to the number of observations per student. The estimates of teacher effectiveness, $\hat{\mathbf{T}}$, are consistent estimators of $\mathbf{T}$ if the number of student test-score observations per teacher goes to infinity as $N \to \infty$. This would be achieved, for example, if each

15. In principle I could update the teacher effects grade by grade, eliminating the issue that the same teacher will appear in both the left and right hand side of Equation 13. However, in practice I have never had convergence issues when updating in one step.

entering cohort of students experienced the same teacher population. In this case, consistency of $\hat{\mathbf{T}}$ would then imply consistency of $\hat{\delta}$. In practice, however, teachers move in and out of the profession, making it likely that as the number of students increases, so do the number of teachers. As a result, the teacher effects themselves will be inconsistent, as will $\hat{\delta}$. However, the Monte Carlo exercises in Section IIIC illustrate that the bias in the estimate of teacher persistence is negligible when the median number of student observations per teacher is only twenty.

The fact that $\hat{\mathbf{T}}$ provides a noisy measure of true teacher value added has important implications for the estimate of dispersion in teacher quality. Simply using the variance of $\hat{\mathbf{T}}$ as an estimate of $\hat{\sigma}^2_T$ will lead to an overstatement of the variance in teacher quality since $\hat{\mathbf{T}}$ contains sampling error, as shown below

$$\hat{T}_{jg} = T_{jg} + \nu_{jg} \tag{14}$$

where ν_{jg} is the sampling error for the jth teacher in grade g. Assuming that ν_{jg} is uncorrelated with T_{jg}, the Var($\hat{\mathbf{T}}$) is equal to $\sigma^2_T + \sigma^2_\nu$. A crude way to correct for the sampling bias is to subtract the average sampling variance across all the teacher estimates from the overall variance of the estimated teacher effects.[16] To do this requires estimating the standard errors for all of the individual teacher effects. I accomplish this by bootstrapping the student sample with replacement and reestimating the model. The standard deviation of a teacher effect across the bootstrap samples provides an estimate of the teacher effectiveness standard error. The Monte Carlo exercises to follow show that this approach for recovering the dispersion in teacher quality works quite well.

The final issue I want to address in terms of the baseline model is the issue of exogeneity. At the start of this section I assumed that ε_{ijg} was uncorrelated with unobserved student ability and teacher assignments in all grades, not just grade g. One obvious concern with this assumption is if there exist time-varying student and classroom characteristics that happen to be related to the teacher assignment. In the expanded version of the model, I address this issue by allowing for observable time-varying classroom and student characteristics.[17] Similar to the other models in the literature, one issue I cannot address is the extent to which parents substitute for teacher quality. In other words, if a student is assigned an ineffective teacher, the parents of that student may compensate by substituting their own time.[18] To the extent that this occurs I will understate the overall variation in teacher quality.

Although sorting on observable, time-varying classroom or student attributes is simple to address, more problematic is sorting on lagged test-score outcomes. The

16. In contrast, Kane and Staiger (2008) treat each teacher effect as a random effect. The dispersion in teacher quality is estimated using the correlation in the average classroom residual across classes taught by the same teacher. The authors note that fixed effects and OLS yield very similar estimates of teacher value added in their sample.

17. Unobserved classroom-year level shocks, such as a dog barking on the day of the test, are ruled out in this framework. However, there existence would likely bias the persistence effect downward and the dispersion in teacher quality upward. The biases will depend on the variance of the classroom-year level shocks and on the extent to which teachers are observed across multiple years.

18. Todd and Wolpin (2007), 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.

assumption that ε_{ijg} is uncorrelated with all teacher assignments in grade $g' > g$ rules out any sorting directly on A_{ijg}. In other words, all the sorting into classrooms has to be based strictly on α.[19] This is a critical assumption in all value-added papers that employ student and teacher fixed effects in either a levels or growth framework. For the remainder of the paper I also maintain the assumption of conditional random assignment.[20]

B. Model Extensions

1. Heterogenous Persistence

The discussion in the previous section relied on the assumption of a constant geometric rate of persistence. 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. The baseline production function can easily accommodate this by simply indexing δ by grade, as seen below.

$$(15) \quad A_{ijg} = \alpha_i + \sum_{j \in J_g} I_{ijg} T_{jg} + \sum_{g'=1}^{g-1} \delta_{g'} \left(\sum_{j \in J_{g'}} I_{ijg'} T_{jg'} \right) + \varepsilon_{ijg}$$

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 nonlinear least squares. Also, in the first-order conditions for α_i and T_{jg}, 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 persist at a geometric rate. Teacher inputs may decay very quickly after one year, but then reach a steady state where the effects no longer decline. This would imply that teachers early in the education process have a significant long term effect on achievement growth. Schools could use this information to find the optimal teacher allocation. The production function would now take the following form

$$(16) \quad A_{ijg} = \alpha_i + \sum_{j \in J_g} I_{ijg} T_{jg} + \sum_{g'=1}^{g-1} \delta_{g-g'} \left(\sum_{j \in J_{g'}} I_{ijg'} T_{jg'} \right) + \varepsilon_{ijg}$$

19. Rothstein (2010) develops a test for whether classroom assignments are random conditional on unobserved student ability and finds that for a cohort of North Carolina fifth graders, classroom assignments are not conditionally random. Using a different sample, Koedel and Betts (2010) show that when multiple cohorts of student test scores are utilized, conditional random assignment cannot be rejected. In addition, Kinsler (Forthcoming) shows that the proposed test in Rothstein (2010) is incorrectly sized for samples similar in size to the one he employs.

20. Sorting on lag test scores combined with persistence in ε_{ijg} will certainly lead to biased estimates of teacher quality in the proposed model. However, it is possible to obtain unbiased estimates of teacher quality in this scenario if students are sorted based on lag test scores and there is no permanent unobserved student heterogeneity. A lag-score model that includes not only the contemporaneous teacher but also controls for the lagged teacher will yield unbiased estimates of the contemporaneous teacher's value-added. Additional details available upon request.

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, Barrow, and Sander (2007). In a lag score framework, not only is it critical that all inputs persist at the same rate, including teacher, school, and student ability, but that the persistence 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.[21]

2. Time-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 appears 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. Other examples of time-varying teacher characteristics include attainment of a graduate degree or licensure status.

The achievement production function can easily accommodate changes to teacher effectiveness that result from variation in observable teacher characteristics over time. Consider the following production function for student achievement

$$(17) \quad A_{ijgt} = \alpha_i + \sum_{j \in J_g} I_{ijgt}(T_{jg} + \beta_2 X_{T_{jgt}}) + \sum_{g'=1}^{g-1} (\delta^{g-g'}(\sum_{j \in J_{g'}} I_{ijgt'}(T_{jg'} + \beta_2 X_{T_{jg't'}}))) + \varepsilon_{ijgt}$$

where $X_{T_{jgt}}$ are the observable characteristics of teacher T_{jg} at time t, and the indicator function is now also indexed by time. Notice that the effect of the observable teacher components enters contemporaneously and in the measure of lagged teacher effectiveness. This allows, for example, the long-run effect of a teacher to vary according to when a student is matched with that teacher. The interpretation of the unobserved teacher value-added estimates is now the expected long-term teacher effectiveness once sufficient experience, education, or licensure is obtained.

Estimation of the model with time-varying teacher characteristics continues to follow the same three steps outlined for the baseline framework. Step 1 needs to be expanded to include not just estimation of δ, but also estimation of β. The first-order conditions required for updating the student and teacher effects in Steps 2 and 3 are altered slightly to account for the time-varying teacher characteristics.

Similar to teachers, students and classrooms will vary over time in observable ways. For example, students may switch schools or repeat a grade. Both of these are likely to impact performance in a particular year. Also, class size and composition will likely vary over time for each student. Incorporating these observable attributes into the above framework follows in the same fashion as the time-varying teacher attributes. Lagged observable student or classroom characteristics can be included in the production function with their own rates of persistence.

21. See Harris and Sass (2008) and Todd and Wolpin (2007) for further discussion of the lag score model.

C. Monte Carlo Evidence

As noted in Section III.A, the persistence and teacher value-added parameters are only consistent under the assumption that as the number of students grow, the population of teachers is held fixed. In reality this is unlikely to hold, as teachers tend to move in and out of the profession often. To provide some small sample evidence regarding the performance of the baseline estimator and the simple extensions outlined in the previous sections I conduct some Monte Carlo experiments. The structure of the data used for the Monte Carlo exercises is chosen to mimic as closely as possible the North Carolina primary school data that will eventually be used to estimate the model. For each model specification, baseline, heterogenous persistence, and time-varying teacher quality, I assume that the methodology for assigning students and teachers to schools and classrooms is identical. Differences only emerge when generating student outcomes since this will depend on the particular specification employed.

The basic structure of the data is as follows. I create seven cohorts of students, each containing 1,500 students. Within each cohort, students are sorted into 25 schools according to their unobserved ability, which is distributed in the population $N(0,.85^2)$. The ratio of the average within school standard deviation in ability to the population standard deviation in ability is approximately 0.92. For each student, I assume a pretest score is available, followed by third, fourth, and fifth grade outcomes. The third through fifth grade classrooms contain 20 students each, and there are three classrooms per grade per school.

Teachers are also sorted into schools based on their unobserved quality, which is distributed in the population $N(0,.25^2)$. The ratio of the average within school standard deviation in teacher quality to the population standard deviation in teacher quality is approximately 0.87. I assume that teachers and students are sorted independently. In other words, while the best teachers tend to end up in the same school, they do not necessarily end up in the school with the best students. With each new cohort of students I assume that there is significant turnover in the teacher population. Turnover is independent of teacher ability, but not experience. This allows me to generate a skewed distribution for the number of observations per teacher, where the mean number of observations per teacher is approximately equal to 35 and the median is only 20. I maintain the sorting on teacher quality across schools even as the teachers within each school change over time.

The final component that remains is how to assign students and teachers to classrooms within each school/grade combination. I assume that teachers and students are negatively sorted on unobserved ability and quality such that the overall correlation between student ability and teacher quality is approximately -0.05. I chose this to match the results obtained with my estimation sample. Note that both the negative correlation between student and teacher ability at the classroom level and the sorting of teachers into schools tends to generate positive correlation in teacher quality across grades. Recall that this is the source of the bias in teacher value-added estimates generated by the levels and growth models.

Once all the matches between students and teachers have been created, I generate test-score outcomes 250 times according to either the baseline accumulation function or one of the extensions previously discussed. In the baseline model, I assume that teacher inputs persist at a rate equal to 0.35. I assume that the test-score measurement

error is independently and identically distributed across grades according to $N(0,.45^2)$. Finally, in an effort to mimic actual data conditions, I assume that some of the student test scores are missing. If a student's fifth grade score is missing, for example, then this student cannot help identify the effectiveness of the assigned fifth grade teacher. However, if a student's third grade score is missing, this student still aids in identifying the third grade teacher effect since the third grade teacher continues to impact that student's test-score outcomes in Grades 4 and 5.

Table 2 shows the results of the Monte Carlo experiments for the baseline model, nongeometric decay, and time-varying teacher quality models. The first column of results show that in the baseline model, the teacher persistence parameter and the dispersion in teacher quality are estimated quite precisely. The remaining columns of Table 2 provide evidence that the extensions to the baseline model, nongeometric decay and time-varying teacher quality, do not inhibit the performance of the proposed estimator. For the model with time-varying teacher quality I assume that teacher experience falls into one of three categories, less than two years, between two and five years, and more than five years of experience. I also allow for a time-varying student attribute, whether the student has transferred schools. The point estimates are generally close to the truth and precisely estimated. The Monte Carlo evidence indicates that the proposed estimators work quite well when the median number of student observations per teacher is equal to 20. I now proceed to discuss the actual schooling data employed in my analysis.

IV. Data

I estimate the cumulative production function detailed in the previous section using administrative data on public school students in North Carolina made available by the North Carolina Education Research Data Center. The data contain the universe of public school students, teachers, and schools across the state. I focus on eight student cohorts who attended third grade between 1998 and 2005. The basic information available for each cohort include 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.

In order to isolate individual measures of teacher effectiveness, the ability to link student outcomes with individual teachers is imperative. Therefore, I use only student test-score observations from self-contained classrooms in Grades 3 through 5.[22] Classrooms are identified by a unique teacher id, however, not all classrooms are self-contained.[23] Using the teacher identifiers, however, I am able to link to class and teacher specific information that allows me to determine whether a class is self-contained.

Each year, there are a significant number of teacher identifiers that cannot be linked to further teacher and classroom information. In order to avoid eliminating a

22. Starting in sixth grade, most students begin to switch classrooms throughout the day. In Grades 6 through 8 students still take one math and reading exam at the end of the year, making it difficult to isolate the impact of each teacher.

23. Student observations associated with charter schools are eliminated from the analysis since there is no information available about teachers that would allow me to determine whether classrooms are self-contained.

Table 2
Monte Carlo Evidence for Various Accumulation Models

	Baseline	Nongeometric Persistence	Varying Teacher Quality
Teacher Persistence	0.349		0.350
	(0.024)		(0.022)
	[0.35]		[0.35]
One-period persistence		0.351	
		(0.024)	
		[0.35]	
Two-period persistence		0.039	
		(0.039)	
		[0.05]	
Less than two years experience			−0.251
			(0.026)
			[−0.25]
Between 3 and 5 years experience			−0.101
			(0.017)
			[−0.1]
Lag student transfer			0.228
			(0.158)
			[0.25]
Student transfer			0.151
			(0.016)
			[0.15]
σ_T^2, Unadjusted	0.076	0.080	0.077
σ_T^2, Adjusted	0.062	0.065	0.062
	[0.062]	[0.065]	[0.062]
R-square	0.86	0.85	0.85
Test-score observations	35,604	35,709	35,616
Students	10,500	10,500	10,500
Teachers	839	875	813

Note: Results are averages across 250 simulations. Standard deviations across the simulations are included in parentheses. True parameter values are included in brackets. σ_T^2 is the estimated variation in teacher quality across all grades. Data generation for the Monte Carlos is described in detail in Section IIIC. The three panels of results reflect three different underlying data generating processes. All models include unobserved student ability and unobserved teacher ability in addition to the parameters listed in the table. Across all three models, the average, median, and minimum number of student observations per teacher are approximately 31, 20, and 9. Estimation follows the procedures discussed in Section IIIB.

large chunk of the data, I assume that if the other classrooms in the same school and grade are self-contained, then the classrooms that cannot be linked are also self-contained. The teacher identifiers that cannot be linked within a particular year cannot be linked across years. Thus, for these teachers I will have only one year of

student observations. The fact that I cannot link these teachers to any observable characteristics, such as experience or education, does not pose a problem for models that allow for time-varying teacher quality since they are observed at only one point in time and do not aid in identifying the effects of the time-varying teacher attributes.

Beyond limiting the sample to students in Grades 3 through 5, I try to minimize as much as possible any other sample restrictions. One benefit of the cumulative model outlined in the previous section is that it does not require balanced student panels, nor does it require students to remain in the same school over time. As a result, I am able to include observations from students with missing test scores, students who eventually leave to attend charters, or switch to schools with classrooms that are not self-contained. For a model focused on estimating teacher quality, incorporating these students is critical since they have outcomes that are significantly different from students who progress from third to fifth grade without ever missing a test or switching schools. However, some data cleaning is employed to minimize coding errors. Classes with fewer than five students or greater than 35 students are excluded from the sample, as are classes with students from more than one grade. Students who skip a grade or repeat a grade more than once are excluded. When imposing all of these restrictions, I try to retain any valid student observations that can be used for the model. As an example, suppose a student attends a class in fifth grade that has fewer than five students, but has valid scores and teacher assignments in third and fourth grade. This student remains in the sample and aids in identifying the value-added of the third and fourth grade teachers.

The final data cleaning step is to ensure that each teacher has a minimum number of associated test-score observations. As noted earlier, the accuracy of the each teacher quality estimate will depend on the number of test-score observations available. Thus, to ensure that there is information in each teacher effect, I include only those teachers with at least ten student test-score observations. To impose this restriction, I have to toggle between eliminating teachers and students since by eliminating one teacher, I will likely invalidate a set of student observations.[24] However, after a few iterations I reach a sample that satisfies the teacher restrictions. I create separate samples, one for math test-score outcomes, and one for reading test-score outcomes. After imposing all the above restrictions, I am left with a sample of approximately 700,000 students, 40,000 teachers, and 2.5 million test-score observations. Close to three-quarters of the entire universe of students who ever attend third through fifth grade between 1998 and 2007 are included in the sample. Notice that students average more than three test-score observations across Grades 3 though 5. This reflects the fact that at the start of Grade 3, students take a pretest in both math and reading. I treat this pretest as an unbiased measure of student ability that is unaffected by prior teacher or other school inputs.[25]

24. After eliminating a set of teachers, students with fewer than two observations and students who have incomplete teacher histories are eliminated. An incomplete teacher history is problematic for estimating the impact of the contemporaneous teacher since it is not possible to account for the influence of all previous teachers. Note that if a student is simply missing a past test score this is not a problem since I can still account for the lasting impact of the teacher associated with the missing score.

25. Clearly to the extent that the kindergarten, first, and second grade teachers persist, part of the ability measure is likely not permanent. However, I have no way of assigning students to classrooms or even schools prior to third grade. In schools with excellent K-2 teachers, unobserved student ability will likely be overstated, leading to downward bias in teacher quality.

Table 3
Summary Statistics: North Carolina Elementary Students and Teacher

	Math	Reading
Teacher statistics		
Total teachers	38,782	38,757
Average observations per teacher	63.75	63.6
Median observations per teacher	24	24
Teacher experience	12.73	12.73
Less than 5 years of experience	0.32	0.32
Graduate degree	0.26	0.26
Student statistics		
Total students	689,641	687,445
Average observations per student	3.58	3.58
Nonwhite	0.40	0.39
Class size	22.9	22.9
Repeat	0.01	0.01
Transfer	0.04	0.04
Missing scores	0.003	0.003
Pretest score	0.12	0.12
	(0.97)	(0.99)
Grade 3 Test Score	0.17	0.02
	(0.94)	(0.99)
Grade 4 Test Score	0.2	0.04
	(0.97)	(0.98)
Grade 5 Test Score	0.18	0.07
	(0.097)	(0.91)

Note: Sample is constructed using cohorts of North Carolina third grade students who enter between 1998 and 2005. Sample selection is discussed in Section IV. Overall, close to three-quarters of the entire universe of students who ever attend third through fifth grade between 1998 and 2007 are included. Math and reading end-of-grade exams are available at the end of third, fourth, and fifth grade. In addition, a pretest score is available from the beginning of third grade. Scores are normalized using the means and standard deviations of test scores in standard setting years as suggested by the North Carolina Department of Instruction. Observations per Teacher indicate the number of student test scores associated with a particular teacher. Graduate degree is an indicator that a teacher received any advanced degree. Repeat is an indicator that a student is repeating the current grade. Transfers indicate that the student is new to the current school.

Table 3 provides some basic information about the math and reading samples employed in estimation. Both samples contain close to 40,000 teachers. Teachers are observed with an average of 64 students, however, the median number of observations is significantly smaller at 24. This reflects both significant churning in the teaching profession early in a teacher's career, as well as the fact that some teacher i.d.s cannot be matched, ensuring that I only observe them for one year. For matched teachers, the average experience level is 13 years, but almost a third of the sample has fewer than five years of experience. As for students, the sample is evenly split among males and females, with nonwhite students making up approximately 40

percent of the sample. There are very few transfers, missing scores, or repeaters. Close to 67 percent of the sample progresses through Grades 3 though 5 without missing any test scores, moving out of the state, switching to a charter school, or switching to a school without self-contained classrooms.

The test scores in third, fourth, and fifth grade come from state mandated end-of-grade exams. The scores are normalized using the means and standard deviations of test scores in standard setting years as suggested by the North Carolina Department of Instruction.[26] Because both the math and reading exams changed during the time frame, the normalizations vary according to the year of the exam. As an example, the first standards setting year is 1997 for both reading and math. In 2001, a new math exam was put in place, making 2001 the new standard. Thus from 1997–2000, all test scores are normalized using the means and standard deviations in 1997. This allows for test scores to improve over time either because students or teachers are improving. Ideally, the method should be employed using a vertically scaled test that remains consistent over time. Overall, the mean test scores are slightly positive, suggesting that even with the limited amount of data cleaning students are still positively selected. The fact that student performance appears to have improved over time also results in positive test-score means.

As Section II illustrates, the amount of teacher and student sorting across schools and classrooms can have important implications for estimates of teacher quality. To provide a sense for the type of sorting in North Carolina's primary schools, I examine the dispersion in student test scores at the population, school, and teacher level in Table 4. In addition to examining the dispersion in contemporaneous scores, I also look at the dispersion in lag scores and pretest scores based on the third, fourth, and fifth grade classroom assignments.

Table 4 illustrates that there exists significant sorting at the school and classroom level, regardless of whether the contemporaneous, lag, or pretest score is considered. The fact that contemporaneous outcomes vary less and less as we move from the population to the classroom suggests that students are sorted by ability into schools and classes, and that teacher quality likely varies significantly across classrooms. The model with student and teacher effects will help disentangle these two components. The data is also generally consistent with the notion that there is limited sorting on lagged student outcomes. Note that the ratio of the within teacher variation to the within school variation in lag scores is very similar across third, fourth, and fifth grade. However, the lag score in third grade, which is actually the pretest score, is not observed at the time third grade teachers are assigned. As a result, the sorting into third grade teacher assignments likely reflects sorting on unobserved student ability. In addition, the magnitude of the within teacher sorting based on pretest scores in fourth and fifth grade is quite similar to the sorting based on lag-scores. This result is not consistent with a process that has principals assigning teachers based strictly on test-score outcomes in the previous grade.

V. Results

Using the North Carolina public school data, I estimate multiple versions of the cumulative model of student achievement. I start with the baseline

26. http://www.dpi.state.nc.us/accountability/reporting/

Table 4
Evidence of Student and Teacher Sorting in NC

	Math			Reading		
	Population	Within-School	Within-Teacher	Population	Within-School	Within-Teacher
Third Grade School/Teacher Assignments						
Standard deviation third grade score	0.913	0.857	0.817	0.957	0.907	0.881
Standard deviation lag score	0.952	0.897	0.868	0.980	0.937	0.910
Fourth Grade School/Teacher Assignments						
Standard deviation fourth grade score	0.959	0.898	0.851	0.960	0.905	0.878
Standard deviation lag score	0.904	0.847	0.815	0.944	0.893	0.868
Standard deviation pretest score	0.945	0.891	0.857	0.975	0.932	0.901
Fifth Grade School/Teacher Assignments						
Standard deviation fifth grade score	0.961	0.894	0.853	0.887	0.834	0.807
Standard deviation lag score	0.953	0.894	0.858	0.950	0.895	0.870
Standard deviation pretest score	0.942	0.889	0.854	0.972	0.930	0.898

Note: Sample is constructed using cohorts of North Carolina third grade students who enter between 1998 and 2005. Sample selection is discussed in Section IV. Overall, close to three-quarters of the entire universe of students who ever attend third through fifth grade between 1998 and 2007 are included. Math and reading end-of-grade exams are available at the end of third, fourth, and fifth grade. In addition, a pretest score is available from the beginning of third grade. Scores are normalized using the means and standard deviations of test scores in standard setting years as suggested by the North Carolina Department of Instruction. The standard deviations within each section are taken with respect to the group identified in the heading. Note that the population standard deviation of the third grade score does not equal exactly the population standard deviation of the lag score using the fourth grade assignment. This simply reflects that not every individual observed in third grade is also observed in fourth grade.

version which assumes a constant geometric decay rate and time-invariant teacher and student ability. The results from this simple accumulation model are then contrasted with the results obtained utilizing the levels and growth frameworks. Finally, I estimate more flexible accumulation models that provide greater insight into the true underlying production of student achievement.

A. Baseline

Table 5 presents estimates from the baseline accumulation model, the levels specification, and the growth specification for math and reading scores. The results indicate that in both math and reading, teacher effects persists at a rate that is neither 0 or 1. Teacher effects persist at a rate equal to 0.38 for math outcomes and 0.32 for reading outcomes. These results fall in the range of previous estimates of the decay rate discussed in the introduction.

The overall dispersion in teacher quality is estimated to be quite significant for math outcomes, and somewhat smaller for reading outcomes. A one standard deviation increase in contemporaneous teacher quality increases math (reading) test scores by approximately 0.25 (0.14) of a standard deviation of the test-score distribution. These results are larger than previous estimates, partly because I do not include school effects.[27] Including school effects in the baseline model indicates that a one standard deviation increase in teacher quality increases math scores by only 0.2 of a standard deviation of the test-score distribution. This effect is identified only from variation in teacher quality within each school-grade combination. The persistence parameter is unchanged when school effects are incorporated.

The final two columns in Table 5 provide estimates of the dispersion in teacher quality under the assumptions that teacher inputs do not persist at all, or perfectly persist. They are included to illustrate the importance of modeling student achievement as a cumulative process. For math outcomes, the overall importance of teacher quality is understated in the levels framework and overstated in the growth framework. For reading outcomes, the levels and growth models result in upward-biased estimates of teacher dispersion, however, the differences are quite small. As a result, in the following discussion I focus on the results for math outcomes only.

The result that the overall dispersion in teacher quality is biased downward in the levels model and biased upward in the growth model is a result of the fact that teacher quality is only slightly positively correlated across grades. The unadjusted cross-grade correlation in teacher quality is approximately 0.11, with a slightly smaller correlation between third and fourth grade teachers, and a slightly larger correlation between fourth and fifth grades. Table 1 illustrates that for very low

27. As noted earlier, unobserved classroom-year level shocks, such as a dog barking on the day of the test or a common illness among students in the class will likely bias the persistence effect downward and the dispersion in teacher quality upward. The extent of the bias will depend on the variance of the classroom-year level shocks. To get a sense for how important classroom-year shocks are, I estimate the standard deviation of the classroom-year shocks using the within-class covariance in test score residuals across math and reading exams. When constructing this covariance I exclude classrooms where the associated teacher is observed for fewer than five years. I find that the standard deviation of the classroom level shocks is only 0.044. In additional Monte Carlo exercises I find that including a class-year shock with this standard deviation has a negligible effect on the estimates of the persistence parameter and the dispersion in teacher quality. Results available upon request.

Table 5
Estimates of Baseline Accumulation, Levels, and Growth Models Using NC Data

	Accumulation	Levels Model	Growth Model
Math Outcomes			
Teacher Persistence	0.375	0	1
	(0.007)	—	—
σ^2_T, Unadjusted	0.0676	0.0635	0.0726
σ^2_T, Adjusted	0.0570	0.0547	0.0607
$\sigma^2_{T_{j3}}$, Adjusted	0.0546	0.0459	0.0550
$\sigma^2_{T_{j4}}$, Adjusted	0.0628	0.0577	0.0648
$\sigma^2_{T_{j5}}$, Adjusted	0.0507	0.0601	0.0536
Observations	2,472,252	2,472,252	1,772,197
R-square	0.86	0.86	0.14
Adjusted *R*-square	0.80	0.80	0.12
Students	689,641	689,641	688,573
Teachers	38,782	38,782	38,567
Reading Outcomes			
Teacher Persistence	0.317	0	1
	0.016	—	—
σ^2_T, Unadjusted	0.0344	0.0322	0.0380
σ^2_T, Adjusted	0.0204	0.0207	0.0219
$\sigma^2_{T_{j3}}$, Adjusted	0.0288	0.0233	0.0305
$\sigma^2_{T_{j4}}$, Adjusted	0.0180	0.0196	0.0157
$\sigma^2_{T_{j5}}$, Adjusted	0.0125	0.0187	0.0112
Observations	2,463,093	2,463,093	1,762,790
R-square	0.83	0.83	0.06
Adjusted *R*-square	0.76	0.76	0.04
Students	687,445	687,445	686,055
Teachers	38,757	38,757	38,544

Note: Sample is constructed using cohorts of North Carolina third grade students who enter between 1998 and 2005. Sample selection is discussed in Section IV. Overall, close to three-quarters of the entire universe of students who ever attend third through fifth grade between 1998 and 2007 are included. Math and reading end-of-grade exams are available at the end of third, fourth, and fifth grade. In addition a pretest score is available from the beginning of third grade. Scores are normalized using the means and standard deviations of test scores in standard setting years as suggested by the North Carolina Department of Instruction. The dependent variable is the normalized end of grade math or reading test score. The explanatory variables are unobserved student and teacher ability. In the accumulation model the rate at which past teacher inputs persist is estimated. σ^2_T is the estimated variation in teacher value-added across all grades, while $\sigma^2_{T_{j3}}$, for example, is the estimated variance of teacher value-added among only third grade teachers. Standard errors are obtained through bootstrap.

levels of cross-grade correlation, the levels model will tend to understate the overall dispersion in teacher quality, while the growth model will tend to overstate the overall dispersion in teacher quality.[28] The source of the small positive correlation in teacher ability across grades appears to stem primarily from teacher sorting across schools. The ratio of the average within school standard deviation of teacher quality to the population standard deviation of teacher quality is 0.89, indicating significant sorting of teachers across schools.

Note that in Table 5, I also list separately the estimated within grade teacher variances across the accumulation, levels, and growth models. The cross-grade pattern of bias in the the levels and growth frameworks is also consistent with the empirical example from Section II. For the levels model, the estimated variance of the third and fourth grade teachers is biased downward relative to the accumulation model, while the variance of the fifth grade teacher effects are biased upward. In the growth model the estimated dispersion of teacher quality is extremely similar to the accumulation model for third grade teachers, and is biased upward for fourth and fifth grade teachers.[29] So while the magnitude of the bias in the overall estimate of teacher dispersion isn't terribly large, the within grade biases are significantly greater. For example, the adjusted variance of third grade teachers is 0.055 in the accumulation model and 0.046 in the levels framework. This gap is four times the size of the gap in the estimates of the overall dispersion in teacher quality.

While the biases in the estimated dispersion of teacher quality resulting from the levels and growth models are important, perhaps more important from a policy perspective are the biases in the individual estimates of teacher quality. Many states are currently debating proposals to use value-added estimates of teacher quality to inform important personnel decisions. Thus policymakers, schools, and particularly teachers want to ensure that estimated teacher value-added is accurate. Section II.B discusses the fact that the accuracy of the levels and growth models in estimating teacher value-added will depend on the cross-grade correlation in teacher quality. When the cross-grade correlation in teacher quality is positive, the levels model will yield teacher value-added estimated that are highly correlated with the truth, while the growth model will likely perform significantly worse. The empirical results are consistent with this theoretical finding. The correlation in estimated teacher quality across the accumulation model and the levels model is 0.97, 0.96, and 0.98 for the third, fourth, and fifth grade teachers respectively. The corresponding numbers for the growth model are significantly lower at 0.96, 0.90, and 0.83.

To put the biases in the individual estimates of teacher quality in perspective I perform the following simple thought experiment. Imagine a policy was put in place in 2007 to monetarily reward teachers who perform at an effectiveness level one

28. Note that the results in Table 1 indicate that when the correlation in teacher quality across grades is equal to 0.12, both the levels and growth models tend to understate the overall variation in teacher quality. However, this is the true correlation in teacher quality across grades, while the result from the North Carolina sample is an unadjusted estimate of the cross-grade correlation. In addition, in the Monte Carlo exercises the variance in teacher quality and the cross-grade correlation in teacher quality are constant across grades, neither of which appears to be true in the data.

29. In the growth framework the estimates of teacher quality for third grade teachers should be identical to the estimates from the accumulation model. However, in the sample teachers actually switch grades across years, leading to a very slight upward bias in the dispersion of the third grade teacher effects.

standard deviation above the mean. Using the same data utilized in this paper, a policy maker estimates teacher quality for the entire population of fifth grade teachers in 2007 using either a levels or growth model. The question is how many teachers receive the bonus when they shouldn't and how many fail to receive the bonus when they should? For the levels model, both numbers are small; 1.7 percent of teachers fail to receive the bonus when they should while 1 percent of teachers are unjustly rewarded. If the policy maker relied on the growth model, 2.1 pecent of teachers who should have received a bonus do not, while 7.9 percent of teachers are unjustly rewarded. Overall, close to 10 percent of the 2007 fifth grade teachers would be misclassified in the growth framework.

The results from the baseline model are useful for illustrating the shortcomings of the basic levels and growth models. However, even the baseline accumulation model makes a number of strong assumptions that can be easily relaxed. The next section shifts the focus from contrasting the accumulation model with the basic levels and growth frameworks to analyzing more robust models of student achievement.

B. Heterogenous Persistence and Time-Varying Teacher Quality

The baseline model assumes that teacher inputs persist at a constant geometric rate and that teacher and student ability are time invariant. In this section I relax these assumptions by allowing for more flexible persistence patterns and including time-varying teacher and student characteristics that are informative for student achievement.

The assumption that teacher inputs persist at a constant geometric rate is useful in that the cumulative nature of the production process can be captured by one parameter. However, if the persistence patterns are more nuanced, then not only are the conclusions about the lasting effects of teachers misguided, but the estimated teacher effects themselves will be biased. As a check on this assumption I reestimate the accumulation model allowing for separate short- and long-term persistence rates. The short-term persistence rate is the persistence of teacher inputs after one year, while the long-term rate is then constant for all additional years. The achievement equation takes the following form,

$$(18)\quad A_{ijg} = \alpha_i + \sum_{j \in J_g} I_{ijg} T_{jg} + \delta_1 \Big(\sum_{j \in J_{g-1}} I_{ij(g-1)} T_{j(g-1)} \Big) + \sum_{g'=1}^{g-2} \delta_2 \Big(\sum_{j \in J_{g'}} I_{ijg'} T_{jg'} \Big) + \varepsilon_{ijg}$$

where g ranges from 3 to 5, with a baseline achievement observation in grade 2.

Estimates of the short- and long-term persistence rates for math and reading outcomes are shown in the left hand panels of Tables 6 and 7. In practice, the estimated rates do not differ much from those generated using the estimated geometric rate in the baseline model.[30] At least in this setting, it appears that the constant geometric persistence rate is not a bad assumption, though with a longer panel of achievement outcomes it might be possible to recover more interesting patterns of teacher persistence.

30. The fact that the estimated teacher effects from the baseline model and the model with nongeometric persistence are almost perfectly correlated also suggests that the geometric assumption in this setting is not limiting.

Table 6
Estimates of Extended Accumulation Models in NC: Math Outcomes

Nongeometric Persistence Rates		Varying Teacher Quality	
One period persistence	0.371	Teacher Persistence	0.3254
	(0.006)		(0.021)
Long-term persistence	0.188	No experience	−0.136
	(0.009)		(0.0033)
		One or two years experience	−0.0137
			(0.0032)
		Between 3 and 5 years experience	0.0045
			(0.0021)
		Graduate degree	−0.0268
			(0.0039)
		Class size	−0.0073
			(0.0002)
		Transfer	−0.0235
			(0.0019)
		Repeat	0.677
			(0.0041)
		Lag repeat	0.617
			(0.0051)
		Twice lagged repeat	0.5726
			(0.006)
σ_T^2, Unadjusted	0.0683	σ_T^2, Unadjusted	0.0668
σ_T^2, Adjusted	0.0575	σ_T^2, Adjusted	0.0576
Observations	2,472,252	Observations	2,472,252
R-square	0.86	*R*-square	0.86
Adjusted *R*-square	0.80	Adjusted *R*-square	0.81
Students	689,641	Students	689,641
Teachers	38,782	Teachers	38,782

Note: See previous table notes for sample information and Sections IIIB1 and IIIB2 for model details. σ_T^2 is the estimated variation in teacher value-added across all grades. Standard errors obtained by bootstrap.

While the assumption that teacher effects persist at a constant geometric rate has little effect on the overall results, the same cannot be said for the time-invariant teacher and student ability assumptions. The right hand panels of Tables 6 and 7 present estimates of the student achievement production function that allows for teacher ability to vary according to experience level and education, student ability to vary with observable characteristics such as whether the student transferred schools or is repeating a grade, and time-varying classroom attributes. Under the assumption of a constant geometric rate of persistence, the achievement equation now takes the following form

Table 7
Estimates of Extended Accumulation Models in NC: Reading Outcomes

Nongeometric Persistence Rates		Varying Teacher Quality	
One-period persistence	0.293	Teacher persistence	0.4317
	(0.013)		(0.019)
Long-term persistence	0.077	No experience	−0.1132
	(0.018)		(0.0052)
		One or two years experience	−0.0682
			(0.0031)
		Between 3 and 5 years experience	−0.0332
			(0.0029)
		Graduate degree	0.0346
			(0.0037)
		Class size	−0.0049
			(0.0001)
		Transfer	−0.0228
			(0.0026)
		Repeat	0.5729
			(0.0049)
		Lag repeat	0.5246
			(0.0053)
		Twice lagged repeat	0.05467
			(0.0057)
σ_T^2, Unadjusted	0.0342	σ_T^2, Unadjusted	0.0344
σ_T^2, Adjusted	0.02	σ_T^2, Adjusted	0.0207
Observations	2,463,093	Observations	2,463,093
R-square	0.83	*R*-square	0.83
Adjusted *R*-square	0.76	Adjusted *R*-square	0.76
Students	687,445	Students	687,445
Teachers	38,757	Teachers	38,757

Note: See previous table notes for sample information and Sections IIIB1 and IIIB2 for model details. σ_T^2 is the estimated variation in teacher value-added across all grades. Standard errors obtained by bootstrap.

$$A_{ijgt} = \alpha_i + \beta_1 X_{igt} + \sum_{j \in J_g} I_{ijgt}(T_{jg} + \beta_2 X_{T_{jgt}}) + \sum_{g'=1}^{g-1} (\delta^{g-g'} (\sum_{j \in J_{g'}} I_{ijgt'}(T_{jg'} + \beta_2 X_{T_{jg't'}}))) + \varepsilon_{ijgt} \quad (19)$$

where X_{igt} are the time-varying student and classroom attributes and $X_{T_{jgt}}$ are the time-varying teacher attributes. Notice that the lasting effect of a teacher on a particular student now depends on what the ability of the teacher was at the time of the teacher-student match. As written, the above achievement function assumes that

the time-varying student and classroom attributes do not persist from one grade to the next. However, as Tables 6 and 7 indicate, I do allow the impact of some of these observable variables to persist.

When teacher, student, and classroom observable characteristics are added to the model the persistence of teacher inputs declines for math outcomes and increases significantly for reading outcomes. However, they remain in the range of 0.3 to 0.4. The estimates of individual teacher abilities are highly correlated with the estimates from the baseline model, 0.98 for math and 0.95 for reading. These correlations overstate the similarities since there are a large number of teachers observed with only one cohort of students, meaning their ability estimates should be largely unaffected. Note that these teachers do not aid in identifying the effects of time-varying teacher attributes.

The results pertaining to the time-varying teacher attributes, experience and education, are quite similar to previous findings in the literature. For experience, I group teachers into four categories: no experience, one to two years of experience, three to five years of experience, and more than five years of experience. Conditional on unobserved teacher ability, students assigned to a teacher with no experience will score 14 percent and 11 percent of a standard deviation lower in math and reading respectively relative to students assigned to a teacher with more than 5 years of experience. In math, the experience gap closes very quickly and teachers with only one or two years of experience perform only slightly worse than a more seasoned teacher. For reading outcomes, even teachers with between 3 and 5 years of experience continue to under perform relative to more experienced teachers. Teachers who receive a graduate degree tend to have lower math test scores and higher reading test scores conditional on ability. This pattern could easily be explained if most teachers return to school for a higher degree in something other than math. Unfortunately this data is unavailable.

The impact of student and classroom characteristics is largely as expected. Class size negatively impacts student performance conditional on teacher and student ability, however it is not economically important even contemporaneously. Similarly, students who just transferred in to their current school perform worse relative to years when they did not transfer, though again the effect is quite small. Because the contemporaneous effects for these variables were small, I assume these effects do not persist into future years. The one student characteristic that was quite important was an indicator for whether a student was currently repeating a grade, or had repeated a grade previously. These students tend to perform significantly better than they did prior to being held back. This could reflect some type of catchup on the students part, or simply a significant negative shock in the initial test score.

A nice feature of this extended model is that it isolates underlying teacher ability independent of actual teacher experience. Thus, I can examine whether teachers with more experience are positively or negatively selected. A priori the selection could work either way. Highly productive teachers are more likely to receive tenure, but these same teachers may also have high labor market productivity, making them more likely to exit the profession. To examine this question I estimate a simple pooled regression where the dependent variable is an indicator for whether a teacher

exits the teaching profession permanently in any given year.[31] The independent variables are experience, experience squared, grade dummy variables, and teacher effectiveness in math. In this simple framework, teacher ability is significantly negatively related to exit, though the effect is small.[32] For example, a one-standard deviation increase in teacher ability for a third-grade teacher with no experience reduces the probability of exit by 7 percent. However, it appears as if the effect is highly nonlinear. When I include an indicator variable that captures whether a teacher is in the 95th percentile of the teacher ability distribution, the coefficient is positive and significant. Being in the bottom of the ability distribution also leads to a much greater probability of exiting the profession.

VI. Conclusion

In this paper I illustrate the biases in estimates of teacher quality associated with the strong persistence assumptions inherent in the standard levels and growth models of student achievement. There are two key features of the data that determine the size and direction of the bias in each model. First, the closer the true underlying rate of teacher persistence is to zero or one, the more accurate the levels or growth model will be. Second, the bias in the levels and growth models tends to increase as the cross-grade correlation in true teacher value-added increases in absolute value.

To avoid these biases, I develop a tractable model of student achievement that explicitly accounts for both the accumulation of teacher inputs over time and the nonrandom sorting of students to teachers. The estimation methodology is quite flexible both in regards to the types of inputs that can be included and assumptions regarding the persistence process. While unobserved teacher quality is assumed to be time-invariant, teacher effectiveness can change over time as teachers accumulate experience and additional schooling. Time-varying student and classroom attributes can also be included in the model, with their own rates of persistence.

To deal with the computational burden of having to estimate thousands of teacher and student effects, I pursue an iterative estimation strategy that updates portions of the parameter space with each iteration. The updating equations stem from the first-order conditions of the least squares problem defined by the achievement equation. At every step of the process, the sum of squared residuals is reduced, and iterations continue until the parameters converge.

I implement the proposed methodology using data from North Carolina's public primary schools. A key finding of the paper is the result that teacher inputs decay rather quickly. 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. As a result, 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

31. Teachers may leave my sample but remain in the North Carolina public schools. I do not consider this an exit. However, the results are quite similar if I instead define exit as exiting my sample.
32. Results available upon request.

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.

References

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

Arcidiacono, Peter, Gigi Foster, Natalie Goodpaster, and Josh Kinsler. Forthcoming. "Estimating Spillovers in the Classroom using Panel Data." *Quantitative Economics*.

Carrell, Scott and James West. 2010. "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors." *Journal of Political Economy* 118(3):409–32

Hanushek, Eric and Steven Rivkin. 2006. "Teacher Quality." In *Handbook of the Economics of Education, Volume 2*, ed. Eric Hanushek and Finis Welch, 1051–78. Elsevier.

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

Harris, Douglas and Tim Sass. 2008. "Teacher Training, Teacher Quality and Student Achievement." Unpublished.

Jacob, Brian, Lars Lefgren, and David Sims. 2010. "The Persistence of Teacher-Induced Learning Gains." *Journal of Human Resources* 45(4):915–43

Kane, Thomas, Douglas Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." NBER Working Paper No. 14607

Kinsler, Josh. Forthcoming. "Assessing Rothstein's Critique of Teacher Value-Added Models." *Quantitative Economics*.

Koedel, Cory, Julian Betts. 2010. "Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique." *Education Finance and Policy* 6(1):18–42

Lockwood, J.R., Daniel McCaffrey, Louis Mariano, and Clause Setodji. 2007. "Bayesian Methods for Scalable Multivariate Value-Added Assessment." *Journal of Educational and Behavioral Statistics* 32(2):125–50

Rockoff, Jonah. 2004. "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data." *American Economic Review, Papers and Proceedings* 94(2):247–52

Rothstein, Jesse. 2010. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *Quarterly Journal of Economics* 125(1):175–214

Sanders, William, Arnold Saxton, and Sandra Horn. 1997. "The Tennessee Value-Added Assessment System (TVAAS): A Quantitative, Outcomes Based Approach to Educational Assessment." In *Grading Teachers, Grading Schools: Is Student Achievement a Valid Evaluation Measure?* ed. Jason Millman, 137–62. Thousand Oaks, Calif.: Corwin Press, Inc.