CALDER

National
Center *for* Analysis *of* Longitudinal Data *in* Education Research

TRACKING EVERY STUDENT'S LEARNING EVERY YEAR

URBAN INSTITUTE

A program of research by the Urban Institute with Duke University, Stanford University, University of Florida, University of Missouri-Columbia, University of Texas at Dallas, and University of Washington

Value-Added Models and the Measurement of Teacher Productivity

Douglas Harris,
Tim Sass, and
Anastasia Semykina

WORKING PAPER 54 • DECEMBER 2010

Value-Added Models and the Measurement of Teacher Productivity

Douglas N. Harris
University of Wisconsin - Madison

Tim R. Sass
tsass@fsu.edu
Florida State University

Anastasia Semykina
Florida State University

Contents

Acknowledgements

The authors wish to thank the staff of the Florida Department of Education's K-20 Education Data Warehouse for their assistance in obtaining and interpreting the data used in this study. This work is supported by Teacher Quality Research grant R305M040121 from the Institute of Education Sciences, U.S. Department of Education and the National Center for the Analysis of Longitudinal Data in Education Research (CALDER) supported through Grant R305A060018 to the Urban Institute from the Institute of Education Sciences, U.S. Department of Education. The authors are also grateful to Anthony Bryk for useful discussion of this research.

CALDER working papers have not gone through final formal review and should be cited as working papers. They are intended to encourage discussion and suggestions for revision before final publication.

The Urban Institute is a nonprofit, nonpartisan policy research and educational organization that examines the social, economic, and governance problems facing the nation. The views expressed are those of the authors and should not be attributed to the Urban Institute, its trustees, or any of the funders or supporting organizations mentioned herein, including the Florida Department of Education. Any errors are attributable to the authors.

CALDER, The Urban Institute
2100 M Street N.W., Washington, D.C. 20037
202-261-5739 • www.caldercenter.org

Value-Added Models and the Measurement of Teacher Productivity
Douglas Harris, Tim Sass, and Anastasia Semykina
CALDER Working Paper No. 54
December 2010

Abstract

Research on teacher productivity, and recently developed accountability systems for teachers, rely on value-added models to estimate the impact of teachers on student performance. The authors test many of the central assumptions required to derive value-added models from an underlying structural cumulative achievement model and reject nearly all of them. Moreover, they find that teacher value added and other key parameter estimates are highly sensitive to model specification. While estimates from commonly employed value-added models cannot be interpreted as causal teacher effects, employing richer models that impose fewer restrictions may reduce the bias in estimates of teacher productivity.

I. Introduction

In the last dozen years the availability of administrative databases that track individual student achievement over time and link students to their teachers has radically altered how research on education is conducted and has brought fundamental changes to the ways in which educational programs and personnel are evaluated. Until the late 1990s, research on the role of teachers in student learning was limited primarily to cross-sectional analyses of student achievement levels or simple two-period studies of student achievement gains using relatively small samples of students and teachers.[1] The advent of statewide longitudinal databases in Texas, North Carolina and Florida, along with the availability of micro-level longitudinal data from large urban school districts has allowed researchers to track changes in student achievement as students move between teachers and schools over time. This in turn has permitted the use of panel data techniques to account for the influences of prior educational inputs, students and schools when evaluating the contributions of teachers to student achievement.

The availability of student-level panel data is also fundamentally changing school accountability and the measurement of teacher performance. In Tennessee, Dallas, New York City and Washington DC, models of individual student achievement have been used to evaluate individual teacher performance. While the stakes are currently low in these cases, there is growing interest among policymakers to use estimates from student achievement models for high-stakes performance pay, school grades, and other forms of accountability. Chicago, Denver, Houston and Washington DC have all adopted compensation systems for teachers based on student performance. Further, as a result of the federal *Teacher Incentive Fund* and *Race to the Top* initiatives, many more states and districts plan to implement performance pay systems in the near future.

[1] For reviews of the early literature on teacher quality see Wayne and Youngs (2003), Rice (2003), Wilson and Floden (2003) and Wilson, et al. (2001).

Measurement of teacher productivity in both education research and in accountability systems is often based largely on estimates from panel-data models where the individual teacher effects are interpreted as a teacher's contribution to student achievement or teacher value-added. The theoretical underpinning for these analyses is the cumulative achievement model developed by Boardman and Murnane (1979) and Todd and Wolpin (2003). However, the assumptions necessary to derive empirical models from the general structural model are generally unstated and untested.

Four recent studies, Todd and Wolpin (2007), Ding and Lehrer (2007), Andrabi, et al. (2009) and Jacob, Lefgren, and Sims (2010), investigate alternative forms of the cumulative achievement function, emphasizing the impact of historical home and schooling inputs on current achievement. Todd and Wolpin (2007) focus on the effect of family inputs on educational outcomes. Assignment of teachers to students within a school is assumed to be exogenous and only school-level averages of teacher inputs are used in their analysis. Ding and Lehrer (2007) exploit data from the Tennessee class-size experiment where students were randomly assigned to teachers and thus avoid the selection problems associated with measuring teacher productivity. Andrabi, et al. (2009) concentrate on the degree to which student learning persists over time and how assumptions about the persistence of prior educational inputs can influence estimates of the determinants of student achievement. Their empirical analysis is limited to the effects of private schools on achievement in Pakistan and thus does not directly evaluate how model specification impacts estimates of teacher value-added. Finally, Jacob, Lefgren, and Sims (2010) focus on the persistence of teacher effects in student achievement models. Using data from North Carolina, they find that 7 percent or more of teacher-induced learning decays within a year.

Three other papers focus more directly on bias in teacher effects and, particularly, non-random assignment of students to teachers. Rothstein (2010) finds evidence that students are dynamically assigned to teachers in ways inconsistent with the usual strict exogeneity assumption in panel-data

models, suggesting that value-added estimates are biased.[2] Kane and Staiger (2008) compare differences in estimated teacher effects from an experiment in which 78 pairs of teachers were randomly assigned to classrooms to pairwise differences in the pre-experiment value-added estimates of teacher productivity. For all of the value-added models that accounted for prior-year student achievement, they could not reject the null that the estimated within-pair differences in teacher productivity were equivalent to the differences under random assignment. A third recent paper, Guarino, Reckase, and Wooldridge (2010), investigates bias attributable to non-random sorting by generating simulated data under various student grouping and teacher assignment scenarios and then comparing the estimates from alternative specification to the known (generated) teacher effects. The simulation approach has the advantage of producing known "true" teacher effects that can be used to evaluate the estimates from alternative models. The disadvantage, however, is that there is no way to know if the selected data generating processes actually reflect the student-teacher matching mechanisms that occur in real-world data.

In this paper we build on the existing literature in several ways. While most recent studies focus on a single aspect of value-added models, we systematically test all of the key assumptions required to derive value-added models from the underlying structural cumulative achievement model. Second, we go beyond simple hypothesis testing and consider how model specification affects the estimated productivity of teachers. By comparing the similarity of teacher effects across different empirical models, we can evaluate the magnitude of the change in teacher rankings of specific modeling choices, each with differing data and computational costs. Third, rather than conduct simulations to compare teacher effect estimates to a hypothetical standard, we evaluate the relative performance of alternative specifications that can be estimated with real-world data.

[2] Koedel and Betts (2009) find evidence that dynamic sorting of student and teachers to classrooms is transitory and that observing teachers over multiple time periods mitigates the dynamic sorting bias.

We begin our analysis in the next section by considering the general form of cumulative achievement functions and the assumptions which are necessary to generate empirical models that can be practically estimated. In section III we delineate a series of specification tests that can be used to evaluate the assumptions underlying empirical value-added models. Section IV analyzes the measurement of schooling inputs that may influence student achievement, including peers, teachers and school-level variables. Section V discusses our data and in section VI we present our results. In the final section we summarize our findings and consider the implications for future research and for the implementation of accountability systems.

II. The Theoretical Achievement Model and Empirical Value-Added Models

A. General Cumulative Model of Achievement

In order to clearly delineate the empirical models that have been estimated, we begin with a general cumulative model of student achievement in the spirit of Boardman and Murnane (1979) and Todd and Wolpin (2003):

$$A_{it} = A_t\left[\mathbf{X}_i(t), \mathbf{F}_i(t), \mathbf{E}_i(t), \mu_{i0}, \varepsilon_{it}\right] \quad (1)$$

where A_{it} is the achievement level for individual i at the end of their t^{th} year of life, $\mathbf{X}_i(t)$, $\mathbf{F}_i(t)$ and $\mathbf{E}_i(t)$ represent the entire histories of individual, family and school-based educational inputs, respectively. The term μ_{i0} is a composite variable representing time-invariant characteristics an individual is endowed with at birth (such as innate ability), and ε_{it} is an idiosyncratic error.

If we assume that the cumulative achievement function, $A_t[\cdot]$ is linear and additively separable,[3] then we can rewrite the achievement level at grade t as:

$$A_{it} = \sum_{h=1}^{t} [\boldsymbol{\alpha}_{ht}\mathbf{X}_{ih} + \varphi_{ht}\mathbf{F}_{ih} + \boldsymbol{\beta}_{ht}\mathbf{E}_{ih}] + \psi_t \mu_{i0} + \varepsilon_{it} \quad (2)$$

where $\boldsymbol{\alpha}_{ht}$, $\boldsymbol{\varphi}_{ht}$ and $\boldsymbol{\beta}_{ht}$ represent the vectors of (potentially time-varying) weights given to individual, family and school inputs. The impact of the individual-specific time-invariant endowment in period t is given by ψ_t.

B. Cumulative Model with Fixed Family Inputs

Estimation of equation (2) requires data on both current and all prior individual, family and school inputs. However, administrative records contain only limited information on family characteristics and no direct measures of parental inputs.[4] Therefore, it is necessary to assume that family inputs are constant over time and are captured by a student-specific fixed component, ζ_i.[5] However, the marginal effect of these fixed parental inputs on student achievement may vary over time and is represented by κ_t. Thus, the effect of the fixed family input ($\kappa_t\zeta_i$) is analogous to the effect of the student component ($\psi_t\mu_i$) in (1).

[3] Figlio (1999) and Harris (2007) explore the impact of relaxing the assumption of additive separability by estimating a translog education production function.

[4] Typically the only information on family characteristics is the student participation in free/reduced-price lunch programs, a crude and often inaccurate measure of family income. Data in North Carolina also include teacher-reported levels of parental education.

[5] In general, one could consider models with uncorrelated unobserved heterogeneity. However, it is likely that the observed inputs (e.g. teacher and school quality) are correlated with the unobserved student effect, which would lead to biased estimates in a random-effects framework. Therefore, in what follows, we assume that the unobserved heterogeneity may be correlated with the observed inputs and focus on a student/family fixed effect.

The assumption of fixed parental inputs of course implies that the level of inputs selected by families does not vary with the level of school-provided inputs a child receives. For example, it is assumed that parents do not systematically compensate for low-quality schooling inputs by providing tutors or other resources.[6] Similarly, it is assumed that parental inputs are invariant to achievement realizations; parents do not increase their inputs when their son or daughter does poorly in school.

The validity of the assumption that family inputs do not change over time is hard to gauge. 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. They have only limited aggregate measures of schooling inputs (average pupil-teacher ratio and average teacher salary measured at the county or state level) and the coefficients on these variables are typically statistically insignificant, whether or not parental inputs are treated as exogenous. Thus it is hard to know to what extent the assumed invariance of parental inputs may bias the estimated impacts of schooling inputs. It seems reasonable, however, that parents would attempt to compensate for poor school resources and therefore any bias in the estimated impacts of schooling inputs would be toward zero.

If we assume that the marginal effects of the endowment and family inputs are the same in each period, i.e., $\kappa_t = \psi_t$ then we can re-label this effect as ω_t and combine the student and family components so that $\omega_t(\zeta_i+\mu_i)=\omega_t\chi_i$.[7] The achievement equation at grade t then becomes:

$$A_{it} = \sum_{h=1}^{t}\left[\boldsymbol{\alpha}_{ht}\mathbf{X}_{ih} + \boldsymbol{\beta}_{ht}\mathbf{E}_{ih}\right] + \omega_t\chi_i + \varepsilon_{it} \qquad (3)$$

[6] For evidence on the impact of school resources on parental inputs see Houtenville and Conway (2008) and Bonesronning (2004).

[7] Note that the marginal effects of fixed parental and child inputs, ω_t, are the same for all students and thus $\omega_t\,\chi_i$ varies over time in the same manner for all students. If the time-varying marginal effect of the individual/family fixed component were student-specific then the effect of the student-specific component in each time period would be perfectly collinear with observed achievement.

Equation (3) represents our baseline model – the least restrictive specification of the cumulative achievement function that can conceivably be estimated with administrative data. In this very general specification current achievement depends on current and all prior individual time-varying characteristics and school-based inputs as well as the student's (assumed time invariant) family inputs and the fixed individual endowment.

C. Age Invariance of the Cumulative Achievement Function

Our baseline model is grade-specific and thus allows for the possibility that the achievement function varies with the grade level. Maintaining this flexibility carries a heavy computational cost, however. In pooled regressions, separate coefficients must be estimated for each input/grade/time-of-application combination. To make the problem more computationally tractable, it is universally assumed in the empirical literature that while the impact of inputs may decay over time, the cumulative achievement function does not vary with the grade level. In particular, it is assumed that the impact of an input on achievement varies with the time span between the application of the input and measurement of achievement, but is invariant to the grade level at which the input was applied. Thus, for example, having a small kindergarten class has the same effect on achievement at the end of third grade as does having a small class in second grade on fifth-grade achievement. This implies that for *any* t:

$$A_{it} = \sum_{h=1}^{t} \left[\boldsymbol{\alpha}_h \mathbf{X}_{i(t+1-h)} + \boldsymbol{\beta}_h \mathbf{E}_{i(t+1-h)} \right] + \omega_t \chi_i + \varepsilon_{it} \qquad (4)$$

D. Accounting for Past Inputs

Given the burdensome data requirements and computational cost of the full cumulative model, even the age-invariant version of the cumulative achievement model, equation (4), has never been directly

estimated for a large sample of students.[8] Rather, various assumptions have been employed about the rate of decay in past inputs in order to reduce the historical data requirements.[9]

Suppose the marginal impacts of all prior student and school-based inputs decline geometrically with the time between the application of the input and the measurement of achievement at the same rate so that for any given h, $\alpha_{(t+1)-h}=\lambda\alpha_{t-h}$, where λ is a scalar and $0 < \lambda < 1.$. With geometric decay the achievement equation can then be expressed as:

$$A_{it} = \sum_{h=0}^{t-1} \lambda^h \left[\boldsymbol{\alpha}\mathbf{X}_{i(t-h)} + \boldsymbol{\beta}\mathbf{E}_{i(t-h)}\right] + \omega_t \chi_i + \varepsilon_{it} \tag{5}$$

Taking the difference between current achievement and λ times prior achievement yields:

$$\begin{aligned} A_{it} - \lambda A_{it-1} = &\left(\sum_{h=0}^{t-1} \lambda^h \left[\boldsymbol{\alpha}\mathbf{X}_{i(t-h)} + \boldsymbol{\beta}\mathbf{E}_{i(t-h)}\right] + \omega_t \chi_i + \varepsilon_{it} \right) - \\ &\left(\sum_{h=0}^{t-2} \lambda^{h+1} \left[\boldsymbol{\alpha}\mathbf{X}_{i(t-1-h)} + \boldsymbol{\beta}\mathbf{E}_{i(t-1-h)}\right] + \lambda\omega_{t-1} \chi_i + \lambda\varepsilon_{it-1} \right) \end{aligned} \tag{6}$$

Collecting terms, simplifying and adding λA_{it-1} to both sides produces:

$$A_{it} = \boldsymbol{\alpha}\mathbf{X}_{it} + \boldsymbol{\beta}\mathbf{E}_{it} + \lambda A_{it-1} + (\omega_t - \lambda\omega_{t-1})\chi_i + \eta_{it} \tag{7}$$

where $\eta_{it} = \varepsilon_{it} - \lambda\varepsilon_{it-1}$.

[8] Todd and Wolpin (2007) estimate the cumulative achievement model using a sample of approximately 7,000 students from the NLSY79-CS. Although they possess good measures of parental inputs and achievement levels they have only a few general measures of schooling inputs measured at the county or state level.

[9] Decay is often interpreted as students forgetting what they learned, but this is not the only interpretation. As Harris (forthcoming) points out, the content of achievement tests may not be completely hierarchical, in the sense that learning the content of 6th grade tests may not contribute directly to measured achievement on 7th grade tests in the same subject. This would also create decay in our equations whether or not students forget anything.

Thus, given the assumed geometric rate of decay, the current achievement level is a function of contemporaneous student and school-based inputs as well as lagged achievement and an unobserved individual-specific effect. The lagged achievement variable serves as a sufficient statistic for all past time-varying student and schooling inputs, thereby avoiding the need for historical data on teachers, peers and other school-related inputs.

Ordinary least squares (OLS) estimation of equation (7) is problematic. Since η_{it} is a function of the lagged error, ε_{it-1}, the lagged achievement term, A_{it-1}, may be correlated with η_{it} , in which case OLS estimates of equation (7) will be biased. Further, even if η_{it} is serially independent and thus uncorrelated with ε_{it-1}, lagged achievement will be correlated with the unobserved student-specific effect, χ_i, since the student-specific effect impacts achievement in each time period.

E. Decay of the Individual-Specific Effect

One key assumption that differentiates empirical specifications of the cumulative achievement model is the treatment of the impact of the individual-specific effect on achievement, $(\omega_t - \lambda\omega_{t-1})\chi_i$. One approach is to assume that time-invariant student/family inputs decay at the same rate as other inputs, λ. In this case $\omega_t = \lambda\omega_{t-1}$ and the individual-specific effect drop out of the achievement equation:

$$A_{it} = \boldsymbol{\alpha}\mathbf{X}_{it} + \boldsymbol{\beta}\mathbf{E}_{it} + \lambda A_{it-1} + \eta_{it} \tag{8}$$

The lagged test score thus serves as a sufficient statistic for the time-constant student/family inputs as well as for the historical time-varying student and school-based inputs. OLS estimates of equation (8) would be consistent so long as the η_{it} are serially independent. If the η_{it} are not serially independent,

one could use instrumental variable (IV) estimation techniques, employing A_{t-2} and longer lags as instruments for A_{t-1}.[10]

An alternative is to assume that the marginal effect of the individual-specific component is constant over time. Thus $\omega_t = \omega_{t-1}$ and $(\omega_t - \lambda\omega_{t-1}) = (1-\lambda)\omega_{t-1}$. Given this assumption $(1-\lambda)\omega_{t-1}$ is a constant, which can be denoted by ϖ. Current achievement is then:

$$A_{it} = \boldsymbol{\alpha}\mathbf{X}_{it} + \boldsymbol{\beta}\mathbf{E}_{it} + \lambda A_{it-1} + \gamma_i + \eta_{it} \tag{9}$$

where $\gamma_i = \varpi\chi_i$ is an individual student fixed effect. If η_{it} are serially independent, equation (9) can be consistently estimated by first differencing (FD) to remove the individual effect and instrumenting for ΔA_{it-1} using A_{t-2} and longer lags as instruments.[11]

F. The Rate of Decay in the Impact of Prior Inputs

A final factor that distinguishes commonly estimated value-added specifications is the assumed rate of decay, λ. There are three typical choices: unrestricted decay, no decay, or complete decay. If no constraints are placed on λ, we are left with either equation (8) or equation (9), depending on whether the fixed student/family inputs are assumed to decay at the same rate as do the time-varying student and school-based inputs or their effect is assumed constant over time ($\omega_t = \omega_{t-1}$). Alternatively, one can assume that there is no decay in the effect of past schooling inputs on current achievement, i.e., (1-

[10] This requires an AR(1) process so that $\mathrm{Cov}(\eta_{it}, \eta_{it-1}) \neq 0$, but $\mathrm{Cov}(\eta_{it}, \eta_{is}) = 0$ for all $|t-s| \geq 2$.

[11] The lagged difference could also be used as an instrument, but it has a disadvantage of imposing restrictions on parameters (in the first-stage regression, where ΔA_{it-1} is the dependent variable, A_{t-2} is forced to have the same coefficient as A_{t-3}). Using A_{t-2} and A_{t-3} instead of the difference is more flexible and is likely to produce higher correlation with the instrumented variable (stronger instruments). So, using lags (rather than differences) is generally preferred.

λ)=0 or λ=1. Given this assumption the coefficient on lagged achievement in equations (8 and 9) is unity.[12] One can then subtract A_{it-1} from both sides of equations (8) and (9) to obtain:

$$\Delta A = A_{it} - A_{it-1} = \boldsymbol{\alpha}\mathbf{X}_{it} + \boldsymbol{\beta}\mathbf{E}_{it} + \eta_{it} \tag{10}$$

$$\Delta A = A_{it} - A_{it-1} = \boldsymbol{\alpha}\mathbf{X}_{it} + \boldsymbol{\beta}\mathbf{E}_{it} + \gamma_i + \eta_{it} \tag{11}$$

Note that in equation (11)) $\gamma_i = \varpi\chi_i = (1\text{-}\lambda)\omega_{t-1}\chi_i$, which is different from zero when the effect of χ_i decays at a rate that is different from that of other inputs (i.e. does not equal one).

As noted by Boardman and Murnane (1979) and Todd and Wolpin (2003), setting λ=1 implies that the effect of each input must be independent of when it is applied. In other words, school inputs each have an immediate one-time impact on achievement that does not decay over time. For example, the quality of a child's kindergarten must have the same impact on their achievement at the end of kindergarten as it does on their achievement in grade 12.

A third alternative is to assume there is no effect of lagged inputs on current achievement. In other words, there is immediate and complete decay so that (1-λ) =1 or λ=0. In this case lagged achievement drops out of the achievement function and equations (8 and 9) become:

$$A_{it} = \boldsymbol{\alpha}\mathbf{X}_{it} + \boldsymbol{\beta}\mathbf{E}_{it} + \eta_{it} \tag{12}$$

$$A_{it} = \boldsymbol{\alpha}\mathbf{X}_{it} + \boldsymbol{\beta}\mathbf{E}_{it} + \gamma_i + \eta_{it} \tag{13}$$

where in equation (12) it is assumed that similarly to other inputs, the impact of the unobserved individual-specific inputs decays immediately (i.e. there is no persistence over time), while in equation

[12] Alternatively, the model can be derived by starting with a model of student learning gains (rather than levels) and assuming that there is no persistence of past schooling inputs on learning gains.

(13) the individual-specific effect is assumed to decay at a different rate (specifically, there is persistence in the unobserved individual–specific inputs).

III. Specification Tests

Most of the assumptions underlying the derivation of popular value-added models, equations (8)-(13), can be tested in a straightforward way, though the data requirements and computational costs can be significant.

A. Testing for Immediate Decay

We can begin by testing the validity of models (12) and (13), which are the most restrictive, but have mild data requirements. These models are correct if there is an immediate decay, i.e. $\lambda = 0$. Model (12) also assumes that there is no unobserved student effect in the original model.

We can test the validity of model (12) by adding lagged inputs and checking whether the lags have significant effects on students' achievement. Under the null, model (12) is correct and the OLS estimator is consistent. Thus, OLS can be used to estimate the augmented equation and usual t-tests or the joint F or Wald test can be used to test the individual or joint significance of lagged inputs.

In model (13), the OLS estimator is consistent only if unobserved heterogeneity (γ_i) is not correlated with the vectors of student and school-based inputs, **X** and **E**. Because in general, observed student, family and school characteristics may be related to unobserved heterogeneity (for example, if more able students are more likely to go to better schools), a more robust approach is to allow arbitrary correlation between γ_i and inputs. In this case, the appropriate estimation method is fixed effects (FE) or first differencing (FD), rather than random effects (RE) or OLS. The immediate decay assumption can again be tested by including lagged inputs in the model and testing their individual or joint significance after the model is estimated by FE or FD.

It is important to note that for FE and FD to be consistent, it is necessary that observed student, family and school inputs are strictly exogenous conditional on the unobserved heterogeneity. In other words, the idiosyncratic shocks to student performance should not be correlated with the choice of inputs in any grade. Rothstein (2010) discusses this problem in detail in relation to estimating teacher effects on student performance. As noted by Rothstein (2010), the strict exogeneity assumption fails if future teacher assignment is partly determined by past and/or current shocks to student performance (for example, if students who experience a drop in their performance are assigned to a class taught by a relatively high productivity teacher next year). In this case, both FE and FD estimators are inconsistent; hence, it is important to check whether strict exogeneity holds. A simple test for strict exogeneity can be performed by adding lead values of inputs in the set of explanatory variables and testing their joint significance in the FE or FD regression (Wooldridge 2002, Chapter 10).[13] If tests reject the null of immediate decay, then testing validity of (12) over (13) is not very useful. Indeed, we have to conclude that both models are wrong.

B. Testing for No Decay

If there is no decay, $\lambda = 1$, so that the appropriate model is the one described by equations (10) and (11). Model (10) follows if $\lambda = 1$ and the impact of unobserved characteristics on achievement is constant over time (i.e., $\omega_t = \omega_{t-1}$), while model (11) assumes that $\lambda \neq 1$ and the unobserved effect is constant over time (i.e., $\omega_t = \omega_{t-1}$).[14] The validity of these models can be checked using the tests that are similar to the ones described above. Specifically, equations (10) and (11) can be augmented by lagged inputs.

[13] This is the same test that Rothstein (2010) uses when testing strict exogeneity in his models VAM1 (regressing test score gains on contemporaneous teacher indicators) and VAM2 (regressing test scores on contemporaneous teacher indicators and the lagged score). Koedel and Betts (2009) use the same test in the model with geometric decay.

[14] Alternatively, model (11) can be derived under the assumption that $\lambda = 1$ and the unobserved effect is trending ($\omega_t = t$).

Under the null hypothesis, model (10) can be consistently estimated by OLS, while model (11) can be consistently estimated by FE or FD.

If the joint F-test or Wald test shows that the lags are jointly insignificant in both models (10) and (11), both models may be appropriate. Insignificance of lagged inputs in the OLS regression suggests that there is no correlation between the unobserved family/student effect and the observed inputs. In this case, the OLS estimation of model (10) is preferred as it is computationally less costly and is consistent under weaker assumptions. However, if the lags are significant in augmented model (10), but not the augmented (11), then model (11) is the one that should be used.

Because the FE and FD estimators are inconsistent if observed inputs are not strictly exogenous, the strict exogeneity assumption has to be tested. Similar to the procedure discussed above, the test can be done by adding the leads of input variables in the FE (FD) regression and subsequently testing their joint significance. If the leads are jointly insignificant, then the hypothesis of strict exogeneity (conditional on the unobserved effect, γ_i) is not rejected, and model (11) can be consistently estimated by FE or FD. If lags are significant in both models, we reject the null of no decay, so that a more complete model has to be considered.

C. Testing for Geometric Decay

If (10) and (11) are rejected than one must consider partial decay models. As discussed above, it is typically assumed that the effects of all educational inputs decay geometrically at some constant rate, λ, so that prior achievement serves as a sufficient statistic for educational inputs received in all previous periods. The corresponding model is given by equation (8) if we assume that the effect of the unobserved heterogeneity decays at the same rate, λ, or, the model is (9) if the unobserved effect is constant over time ($\omega_t = \omega$, $t = 1, \ldots, T$).

Testing for geometric decay is relatively simple in model (8). Assuming that the idiosyncratic error, η_{it}, is not correlated with current inputs and past achievement, OLS is consistent. Also, under the null hypothesis, lagged inputs should not appear in the equation, so the test for geometric decay is again performed by testing the joint significance of lagged inputs. If the error term contains a time-constant unobserved effect, the OLS estimator is biased, and the test is not valid. Instead, one should account for unobserved heterogeneity and focus on model (9).

In model (9), estimation is complicated due to the presence of the lagged dependent variable, which is inevitably correlated with γ_i. Under the standard assumption that the idiosyncratic errors are serially uncorrelated, the common approach is to remove the unobserved effect by first-differencing, and then use second and possibly further lags of the dependent variable to instrument for $\Delta A_{it\text{-}1}$. Using the instrumental variables method is necessary because in the differenced equation, $\text{Cov}(\Delta A_{it\text{-}1}, \Delta\eta_{it}) = \text{Cov}(A_{it\text{-}1} - A_{it\text{-}2}, \eta_{it} - \eta_{it\text{-}1}) = \text{Cov}(A_{it\text{-}1}, \eta_{it\text{-}1})$ because $\text{Cov}(A_{it\text{-}1}, \eta_{it}) = \text{Cov}(A_{it\text{-}2}, \eta_{it\text{-}1}) = \text{Cov}(A_{it\text{-}2}, \eta_{it}) = 0$ when $\{\eta_{it}\}$ are serially uncorrelated. Because $\text{Cov}(A_{it\text{-}1}, \eta_{it\text{-}1}) \neq 0$ by construction, instruments are needed.

Testing for geometric decay in model (9) is performed as follows: (i) augment the model by lagged inputs; (ii) first-difference the equation to remove the unobserved effect and (iii) estimate the differenced equation by two-stage least squares (2SLS) or the generalized method of moments (GMM) with second and possibly further lagged test scores used as instruments. If the geometric decay assumption holds, then the lagged inputs should be jointly insignificant.

To ensure that the described estimator is consistent and the suggested test is valid, it is important to test the validity of the instruments. The assumption of no serial correlation is usually tested by testing H_0: $\text{Corr}(\Delta\eta_{it}, \Delta\eta_{it\text{-}1}) = -0.5$, which is equivalent to testing the assumption that $\{\eta_{it}\}$ are serially independent. In practice, the correlation between the current and lagged residuals in the differenced equation is computed and used for testing. Another standard test is a test that checks whether the instrument (or instruments) is strongly partially correlated with the instrumented variable.

Strict exogeneity of inputs can again be tested by adding lead input values in the equation and testing their joint significance.

If validity of instruments is confirmed and lagged inputs have no significant partial effects in (9), this can be regarded as evidence in support of the geometric decay hypothesis. If the hypothesis is rejected, it may be due to several factors. First, the decay may not be geometric. Second, the decay may be geometric, but the rate of decay is not the same for all inputs. Third, it may be that assuming a constant unobserved student/family effect is too restrictive. Finally, the assumption that the impact of an input on achievement varies only with the time span between the application of the input and measurement of achievement, and does not depend on when it was applied, may fail. These possible explanations can be tested using model (4), as discussed below.

D. Testing Whether the Unobserved Effect is Time-Constant

One way to check whether the unobserved student/family effect is constant over time ($\omega_t = \varpi$, $t = 1, \ldots, T$) involves testing for serial correlation in the first-difference equation. Specifically, assume that the idiosyncratic error in equation (4) follows a random walk, i.e. $\varepsilon_{it} = \varepsilon_{it-1} + e_{it}$, where $\{e_{it}\}$ are not serially correlated and $\text{Cov}(\varepsilon_{it-1}, e_{it}) = 0$. Then, under H_0: $\omega_t = \varpi$, the differenced equation is:

$$\Delta A_{it} = \sum_{h=1}^{t-2}\left[\alpha_h \Delta X_{i(t+1-h)} + \beta_h \Delta E_{i(t+1-h)}\right] + \alpha_1 X_{i1} + \beta_1 E_{i1} + e_{it}\,. \qquad (14)$$

If we make assumptions about ε_{it} as described above, positive serial correlation in the error term in equation (2) can be attributed to the presence of the unobserved effect, which means differencing was not effective in removing unobserved heterogeneity and H_0: $\omega_t = \varpi$ should be rejected.

Another test, which does not rely on the assumption of the idiosyncratic errors following a random walk, is based on the comparison of the FE and FD estimators. If inputs in equation (4) are

strictly exogenous conditional on the unobserved effect, i.e. $E(X_{it}'\varepsilon_{ir}) = 0$ and $E(E_{it}'\varepsilon_{ir}) = 0$ for all t and r [or, $E(\varepsilon_{it} \mid X_{i1}, X_{i2}, \ldots, X_{iT}, E_{i1}, E_{i2}, \ldots, E_{iT}, \gamma_i) = 0$], and $w_t = w$, then both FE and FD estimators of the parameters in equation (4) are consistent and hence, the corresponding estimated effects should not be statistically different. Therefore, a Hausman-type test comparing the FE and FD estimators can be used (Wooldridge 2002). The traditional form of the Hausman test statistic is not applicable unless we make additional assumptions that ensure efficiency of one of the estimators, so the general form of the test statistic is preferred.[15]

Both tests outlined in this subsection maintain the assumption of strict exogeneity (conditional on the unobserved effect) under both the null and alternative. In other words, we have to assume that a shock to student achievement in grade t (ε_{it}) does not affect the choice of inputs in the next grade, conditional on all past inputs and unobserved heterogeneity, γ_i. A familiar test for strict exogeneity can be used to verify the validity of this assumption.

If the time-invariance of the unobserved effect is rejected, a functional form for ω_t has to be specified. A popular and practically feasible choice is to assume that the unobserved effect in equation (4) is trending: $\omega_t = 1 + \xi t$, where t is a time trend (see, for example, Wooldridge 2002). After taking first differences in (4), $\Delta(\chi_i \xi)\, t = \chi_i \xi \equiv \gamma_i$:

$$\Delta A_{it} = \sum_{h=1}^{t-2} \left[\alpha_h \Delta X_{i(t+1-h)} + \beta_h \Delta E_{i(t+1-h)} \right] + \alpha_1 X_{i1} + \beta_1 E_{i1} + \gamma_i + e_{it}\,. \qquad (15)$$

Equation (15) can be consistently estimated by FE or FD, assuming that all inputs are strictly exogenous conditional on the unobserved effect. Of course, α_1 and β_1 cannot be estimated because $\mathbf{X}_{i1}$ and $\mathbf{E}_{i1}$ are time-invariant. Once again, a usual test for strict exogeneity can be used here.

[15] When relative efficiency of one of the two estimators cannot be established, then the asymptotic variance of the difference between the two estimators is not the same as the difference in the two variances; the covariance should be included in the computation of the variance of the difference. See, for example, Wooldridge (2002), Section 14.5.1.

Rothstein (2010) considers a model, where instead of specifying an individual-specific trend, he models the unobserved student-specific component as $\omega_t\chi_i$, where ω_t is common to all students, but may differ by grade (or over time). Rothstein assumes that the impact of observed inputs does not decay (i.e. the score gain is the dependent variable, and only current inputs appear on the right-hand side of the equation) and suggests testing for strict exogeneity by means of Chamberlain's modeling device (Chamberlain 1984).[16] Specifically, the time-constant part of the individual-specific effect is modeled as:

$$\chi_i = \mathbf{X}_{i1}\boldsymbol{\zeta}_1 + \ldots + \mathbf{X}_{iT}\boldsymbol{\zeta}_T + \mathbf{E}_{i1}\boldsymbol{\psi}_1 + \ldots + \mathbf{E}_{iT}\boldsymbol{\psi}_T + \eta_i,$$

where η_i is uncorrelated with the observed inputs by construction. After substituting the whole expression for the unobserved individual-specific component into the model, the reduced-form parameters on the inputs can be consistently estimated by OLS. In each grade, *t*, the effect of the current input on student performance consists of its direct effect, as well as the effect due to non-zero correlation between the input and χ_i. On the other hand, if strict exogeneity holds, then future and past inputs should have non-zero estimated effects on student performance only because of their correlation with χ_i. Because the unobserved student-specific component is given by $\omega_t\chi_i$, only $\omega_t\zeta_1$, …, $\omega_t\zeta_T$, $\omega_t\psi_1$, …, $\omega_t\psi_T$ (t = 1, …, T) can be estimated in the reduced-form equation (that is, it is not possible to estimate ω_t separately from ζ_t or ψ_t).

However, Rothstein (2010) notes that if inputs are strictly exogenous conditional on the unobserved effect, then certain restrictions on parameters should hold. He subsequently uses the reduced-form parameters to recover the initial structural parameters, while also imposing the restrictions implied by the strict exogeneity of inputs. To check whether the strict exogeneity assumption holds, he tests whether the imposed restrictions are valid and concludes that they are not.

[16] Rothstein (2010) only considers teacher-specific effects as inputs. We present the discussion based on our model to avoid introducing new notation.

In what follows, we limit our attention to models with an individual-specific trend and consider a model where the rate of decay is not restricted, as summarized by equation (4).

E. Testing for Input-Specific Geometric Decay

A test for a geometric rate of decay that is the same for all inputs has already been discussed above. However, it would also be useful to know whether the impact of each input (or possibly only some inputs) decays geometrically, even though the rate of decay may not be the same for all inputs. The null hypothesis is

$$H_0: \frac{\alpha_{t-2,j}}{\alpha_{t-1,j}} = \frac{\alpha_{t-3,j}}{\alpha_{t-2,j}} = \ldots = \frac{\alpha_{3,j}}{\alpha_{2,j}},$$

or

$$H_0: \frac{\beta_{t-2,j}}{\beta_{t-1,j}} = \frac{\beta_{t-3,j}}{\beta_{t-2,j}} = \ldots = \frac{\beta_{3,j}}{\beta_{2,j}},$$

for a given input *j*. These are nonlinear hypotheses that can be tested using a Wald-type test. Depending on the outcome of the tests described in the previous subsection, the test should be done after estimation of either equation (4) or equation (15) by FE or FD.

F. Testing for Grade Invariance

Depending on whether the unobserved student/family effect in equation (4) is time-invariant or trending, grade invariance (the assumption that the effect of the input depends only on the time span between the application of the input and measurement of achievement) can be tested following the FE or FD estimation of either equation (4) or (15). Specifically, each input can be interacted with time (or grade) dummies, and the interaction terms can be added to the estimating equation. Under the null

hypothesis of grade invariance, the interaction terms should be jointly insignificant, which can be tested using the usual Wald or F-test.

IV. Measuring the Effects of Teachers and Other Schooling Inputs

A. Decomposing School Inputs

The vector of school-based educational inputs, $\mathbf{E}_{it}$, contains both school-level inputs such as the quality of principals and other administrative staff within school m, $\mathbf{S}_{mt}$, as well as a vector of classroom-level inputs in classroom j, $\mathbf{C}_{jt}$. The vector of classroom inputs can be divided into four components: peer characteristics, $\mathbf{P}_{-ijmt}$ (where the subscript –i students other than individual i in the classroom), time-varying teacher characteristics (such as experience), $\mathbf{T}_{kt}$ (where k indexes teachers), non-teacher classroom-level inputs (such as books, computers, etc.), $\mathbf{Z}_{jt}$, and the primary parameter vector of interest, time-invariant teacher characteristics, δ_k (including, for example, "innate" ability and pre-service education). If we assume that, except for teacher productivity, there is no variation in education inputs across classrooms within a school, the effect of $\mathbf{Z}_{jt}$ becomes part of the school-level input vector, $\mathbf{S}_{mt}$. The baseline value-added model (equation (3)) can then be expressed as:

$$A_{it} = \sum_{h=1}^{t} \left[\boldsymbol{\alpha}_{ht} \mathbf{X}_{ih} + \boldsymbol{\beta}_{1ht} \mathbf{P}_{-ijmh} + \boldsymbol{\beta}_{2ht} \mathbf{T}_{kh} + \delta_{kmh} + \mathbf{S}_{mh} \right] + \omega_t \chi_i + \varepsilon_{it} \tag{16}$$

B. Modeling Teacher Effects

One important specification issue related to teacher effects is whether the impact of teachers on student achievement varies across students and schools. It may be that some teachers possess skills that aid some students more than others or perhaps the racial/ethnic identity of a teacher has differential effects on students of different races and ethnicities. To control for potential variation in

teacher effects among students a number of studies either include interactions between teacher and student characteristics (Wright, Horn, and Sanders, 1997) or conduct separate analyses for different groups of students (Aaronson, Barrow, and Sander 2007; Dee 2004; Goldhaber and Anthony 2007). A recent analysis by Lockwood and McCaffrey (2009) directly investigates whether teacher value added varies across different types of students. Another recent study by Sass, et al. (2010) compares the value added of teachers across different school settings and fails to reject the null that a teacher's effect is the same in different schools.[17]

C. Accounting for Peer Effects and Other Classroom Factors

There is a rapidly growing empirical literature on classroom peer effects. It is well known that if students are assigned to classrooms non-randomly and peer-group composition affects achievement, then failure to control for the characteristics of classroom peers will produce biased estimates of the impact of teachers on student achievement. The measured teacher effects will capture not only the true impact of teachers but will also partially reflect the impacts of omitted peer characteristics. Recognizing this potential problem, the majority of the existing studies of teacher effects contain at least crude measures of observable peer characteristics like the proportion who are eligible for free/reduced-price lunch. An alternative approach, at least in research contexts, is to focus on classes where students are, or appear to be, randomly assigned, as in Clotfelter, Ladd, and Vigdor (2006), Dee (2004), and Nye, Konsantopoulos, and Hedges (2004).

As with the effects of peers, omission of other classroom-level variables can bias the estimated impact of teachers on student achievement if the allocation of non-teacher resources across classrooms is correlated with the assignment of teachers and students to classrooms. For example, principals may

[17] Another issue, which we do not discuss here, is the choice of modeling teacher effects with fixed or random-effects estimators. Lockwood and McCaffrey (2007) provide a detailed comparative analysis of fixed and random effects estimators in the context of student achievement models.

seek to aid inexperienced teachers by giving them additional computer resources. Similarly, classrooms containing a disproportionate share of low-achieving or disruptive students may receive additional resources like teacher aides. Due to the paucity of classroom data on non-teaching personnel and equipment, most studies omit any controls for non-teacher inputs. The only exceptions are Dee (2004) and Nye, Konstantopoulos and Hedges (2004) who use data from the Tennessee class-size experiment where classrooms were explicitly divided into three types: small classes, larger classes with an aide and larger classes with no aide.

D. Measuring School-Level Effects

Typically administrative data provide little information on time-varying school-level inputs like scheduling systems, curricular choices, leadership styles and the like. The lone exception is the identity of the principal and perhaps some basic observable characteristics of the principal like experience or educational attainment. Consequently, it is common to assume that school-level inputs are constant over the time span of analysis and replace the vector of school characteristics, $\mathbf{S}_{mh}$, with a school fixed effect, ϕ_{mh}. When school-level effects are included, the teacher fixed effect measure in equation (16), δ_{kmh}, captures the effect of a given teacher's time-invariant characteristics on her students' achievement relative to other teachers at the same school. Thus, while the school effect controls for unobservable (time invariant) school characteristics and the non-random assignment of teachers to schools, it obviously limits the comparison group for assessing teacher productivity. This is particularly problematic in accountability contexts since one typically wants to compare the performance of a teacher with all other teachers in the school district or state, not just at their own school.

V. Data

In order to test alternative model specifications we utilize data from the Florida Department of Education's K-20 Education Data Warehouse (EDW), an integrated longitudinal database covering all Florida public school students and school employees. Our sample begins with school-year 1999/2000, which is the first year in which statewide standardized testing in consecutive grade levels was conducted in Florida. Our panel continues through the 2003/04 school year.

During our period of analysis the state administered two sets of reading and math tests to all third through tenth graders in Florida. The "Sunshine State Standards" Florida Comprehensive Achievement Test (FCAT-SSS) is a criterion-based exam designed to test for the skills that students are expected to master at each grade level. The second test is the FCAT Norm-Referenced Test (FCAT-NRT), a version of the Stanford-9 achievement test. The Stanford-9 is a vertical or development-scale exam. Hence scores typically increase with the grade level and a one-point increase in the score at one place on the scale is equivalent to a one-point increase anywhere else on the scale. We use FCAT-NRT scale scores in all of the analysis. The vertical scale of the Stanford Achievement Test allows us to compare achievement gains of students with differing initial achievement levels. Further, use of the FCAT-NRT minimizes potential biases associated with "teaching to the test," since all school accountability standards, as well as promotion and graduation criteria in Florida are based on the FCAT-SSS, rather than the FCAT-NRT.

Although achievement test scores are available for both math and reading in grades 3-10, we limit our analysis to mathematics achievement in middle school, grades 6-8. We select middle-school mathematics classes for a number of reasons. First, we require second-lagged scores to serve as potential instruments for lagged achievement. Given that testing begins in grade 3 this precludes analysis of student achievement prior to grade 5.

Second, it is easier to identify the relevant teacher and peer group for middle-school students than for elementary students. The overwhelming majority of middle school students in Florida move between specific classrooms for each subject whereas elementary school students typically receive the majority of their core academic instruction in a "self-contained" classroom. However, for elementary school students enrolled in self-contained classrooms, five percent are also enrolled in a separate math course and nearly 13 percent are enrolled in either special-education or gifted courses.

Third, because middle-school teachers often teach multiple sections of a course during an academic year, it is easier to clearly identify the effects of individual teachers on student achievement. In elementary school, teachers typically are with the same group of students all day long and thus teacher effects can only be identified by observing multiple cohorts of students taught by a given teacher over time. In contrast, both variation in class composition across sections at a point in time as well as variation across cohorts over time help to distinguish teacher effects from other classroom-level factors affecting student achievement in middle school.

Fourth, we choose to avoid high school grades (grades 9 and 10) because of potential mis-alignment between test content and curriculum. At the high-school level math courses become more diverse and specialized. Thus the content of some high school math courses, particularly advanced courses, may have little correlation with concepts being tested on achievement exams.

Finally, we focus on math achievement rather than reading because it is easier to clearly identify the class and teacher most relevant to the material being tested. While some mathematics-related material might be presented in science courses, direct mathematics instruction almost always occurs in math classes. In contrast, middle school students in Florida may be simultaneously enrolled in "language arts" and reading courses, both of which may cover material relevant to reading achievement tests.

These five differences across grades and subjects appear to have practical implications for value-added analysis. Harris and Sass (2009) show that the effects of teacher experience and formal training

programs are much more likely to be positive and statistically significant in middle school math compared with all other grade-subject combinations. They argue, as we do here, that the value-added results for middle school math are probably, though not necessarily, less biased and more precise for the reasons given above.

In addition to selecting middle-school math courses for analysis, we have limited our sample in other ways in an attempt to get the cleanest possible measures of classroom peers and teachers. First, we restrict our analysis of student achievement to students who are enrolled in only a single mathematics course (though all other students enrolled in the course are included in the measurement of peer-group characteristics). Second, to avoid atypical classroom settings and jointly taught classes we consider only courses in which 10-50 students are enrolled. Third, we eliminate any courses in which there is more than one "primary instructor" of record for the class. Finally, we eliminate charter schools from the analysis since they may have differing curricular emphases and student-peer and student-teacher interactions may differ in fundamental ways from traditional public schools.

Estimation of the achievement models with lagged test scores and individual fixed effects requires at least three consecutive years of student achievement data. Given statewide testing began in 1999/2000, our analysis is limited to Florida traditional public school students in grades 6-8 over the years 1999/04 through 2003/04 who took the FCAT-NRT for at least three consecutive years. This includes four cohorts of students, with over 120,000 students in each cohort. Unfortunately, it is not computationally tractable for us to consistently estimate models with lagged dependent variables using the entire sample. We therefore randomly select 25 of Florida's 67 countywide school districts for analysis. Descriptive statistics for the variables in the 25-district data set are provided in Table 1.

VI. Results

A. General Rate of Decay of All Past Inputs

The simplest value-added models assume that the effects of past inputs decay completely and immediately and thus prior educational inputs have no bearing on current achievement. This can easily be tested by determining if prior inputs have significant effects on current achievement. To make the test computationally feasible we limit our analysis to tests of the impact of the prior-year teacher and one, two and three lags of non-teacher schooling inputs. We consider two variants: one where student/family inputs are also subject to immediate decay and one in which student/family inputs are time-constant (captured by student fixed effects). As reported in the first and fourth columns of Table 2, we strongly reject the null that prior inputs have no effect on current achievement in both cases.

The test for no decay (complete persistence) of prior inputs is similar. With no decay, the gain in achievement will be independent of prior inputs. As indicated in Table 2, we strongly reject the null that prior inputs have no effect on current achievement gains for both the model where student/family inputs decay at the same rate as other inputs (second column) and the model where student/family inputs are constant over time (fifth column). This is consistent with the results of Kane and Staiger (2008), which reject the equivalence of value added and experimental teacher effect estimates when value-added models exclude information on prior test scores.

Rejection of the immediate and no-decay models implies there is partial persistence in prior schooling inputs. The simplest form of a partial-decay model is one in which inputs decay at a geometric rate. This allows the prior-year test score to serve as a sufficient statistic for all past schooling inputs. In columns three and six of Table 2 we present tests of the geometric decay model; one assumes that student family inputs decay at the same rate as schooling inputs (and are thus captured by the lagged test score) while the other allows for a time-invariant student/family effect. In both cases we strongly reject the null that the effect of prior inputs on current achievement is zero. This implies that the lagged

test score is not capturing all of the effect of prior inputs on current achievement. The test for geometric decay when student/family inputs are time-constant requires use of instrumental variables (IV). As described above we test for the validity of the instruments by testing whether the correlation between the current and lagged residual is equal to -0.5 and whether the instrument is strongly partially correlated with the instrumented variable. Results for both of these tests confirm the validity of the instruments.

B. Is the Effect of the Unobserved Student/Family Input Time Invariant?

Our rejection of the geometric decay model could be due to a number of factors. For example, a key assumption is that the effects of student/family inputs either decay at the same rate as all other inputs or are time constant. If they decay at the same geometric rate as all other inputs, their effect is captured by the lagged test score and thus they do not directly appear in the model with partial decay. If instead student/family input effects are time-constant, their effect can be swept away through first-differencing or inclusion of fixed effects for students. However, if the effect of the unobserved student/family inputs is not time-constant, this effect will not drop out from the geometric decay model even after differencing, which may lead to the rejection of the model. As described above, the assumption of time-constant unobserved student/family input effects can be tested by looking at the serial correlation in the error terms in the baseline model (equation (4)).

Under the assumption that idiosyncratic errors in that equation follow a random walk, there should be no remaining positive serial correlation in the errors after the equation is estimated on first differences. Thus a test for the time constant assumption can be conducted by testing for serial correlation in the residuals after equation (4) is estimated by FD. Results of such a test are presented in the first column of Table 3. The estimated correlation coefficient between the current and lagged residual is -0.49, where the negative sign indicates that the random walk assumption is likely false. Also,

the fact that the correlation coefficient is so close to -0.5 is consistent with the assumption that in the original model (equation (4)), the unobserved effect is time-constant and the idiosyncratic errors are not serially related.[18] Thus, it appears to be unlikely that the rejection of the geometric decay model is due to the unobserved student/family effect being time-varying.

C. Input-Specific Decay of Past Inputs

Besides non-constant and non-geometric decay of student/family inputs, the geometric-decay model could be rejected if different schooling inputs decay at different rates. For example, the effect of having an experienced third-grade teacher might be longer lasting than the effect of having a small class in third grade. As described above, this can be tested by comparing the ratio of the coefficients on lagged inputs over time, input by input. For computational tractability, we include three lags of each input and test the null that the ratio of the coefficients on the first and second-lagged inputs equals the ratio of the coefficients on the second and third lagged inputs. The results, displayed in Table 3, indicate that we cannot reject the null that each input decays at its own geometric rate. This is true no matter whether we assume that unobserved student/family inputs are subject to immediate decay (OLS model), are time constant (FE, FD models) or follow a time trend (FE on FD model).[19]

D. Tests of Grade-Invariance

Recall that empirical value-added models universally assume that lagged inputs have the same effect on contemporaneous achievement, irrespective of grade level. Thus, for example, grade 6 inputs (class size, peer composition, etc.) have the same effect on 8th grade test scores as do grade 5 inputs on 7th

[18] This finding is also consistent with the results reported in the second and sixth columns in Table 2, where the correlation between the residuals is also reasonably close to -0.5.

[19] In addition to the test results presented in Table 3, we also tested to see if various combinations of inputs share a common decay rate. Occasionally we uncovered cases where we could not reject a common decay rate for two or more inputs, but they were infrequent and did not follow any particular pattern. For example, the effects of various teacher credentials did not decay at similar rates.

grade test scores. This can be tested by interacting each input with time (or grade) dummies, and testing the significance of the interaction terms. Results from estimating such a specification are presented in Table 4. Recall that there are three middle-school grades in the sample, six, seven and eight. Thus we include interactions with grade 6 and with grade 7. We present tests for the joint significance of all grade-input interactions, as well as separate tests for the significance of grade six and grade seven interactions. For all of the first and second-lag interactions we strongly reject the null of grade invariance. Put differently, we do not find support for the common assumption that prior inputs affect achievement in the same way no matter when they are applied.

E. Tests of Strict Exogeneity

In Table 5 we present results from tests of strict exogeneity for several models with varying assumptions regarding the persistence of schooling inputs and the nature of the unobserved student/family input. In every case we strongly reject the null that future teacher assignments have no "effect" on current student achievement. This suggests that student assignment to teachers is based in part on realized prior achievement, which means that value-added estimates of teacher effects will be biased.

F. Differences in Estimates Across Models

In virtually every case we reject the assumptions required to derive commonly estimated value-added models from the structural cumulative achievement model. This implies that either the structural model is inappropriate or (more likely) that empirical value-added models produce biased estimates as a result of imposing invalid assumptions. However, recognizing that value-added models produce biased estimates of teacher productivity does not necessarily invalidate their use. Rather, one must evaluate the magnitude of potential bias against the reliability of alternative forms of teacher evaluation and the associated costs of implementation.

Assuming that potentially biased value-added estimates of teacher productivity are preferred to no estimates at all, one must consider how different models perform relative to each other and what modeling assumptions have the most leverage on the estimates produced by value-added models. For example, relaxing assumptions regarding uniform decay across inputs involves the data collection cost of acquiring information on all observed prior inputs and the computational burden of estimating complex models with many parameters. If the resulting estimates differ little from those generated by more restrictive models, then the slight increase in bias may be worth the savings in data and computation costs.

In Table 6 we present results of tests of the similarity of estimated coefficients for key inputs (free/reduced-price lunch status, disciplinary incidents, class size, student mobility, teacher experience levels, teacher educational attainment, and teacher certifications) as well as for the estimated teacher effects. The comparison tests are t-tests, where the test-statistic is computed as a difference in the coefficient estimates from the two models divided by the absolute value of the difference in standard errors. This test statistic is used for convenience and is computed under the assumption that the estimates in the two models are strongly positively correlated, which results in the largest possible value of the test-statistic.[20] Thus, the test is biased in favor of rejecting the null that the two estimates are the same and underestimates the extent to which different models are comparable. The percentages in Table 6 represent the proportion of coefficients/effects in which we fail to reject the null of equality across models. The comparison models are those with three lags of inputs and no prior test scores.

A number of patterns emerge. First, including lagged test scores (in addition to lagged inputs) has a fairly dramatic effect on the estimated teacher effects, suggesting that prior test scores are picking up the influence of unmeasured prior inputs. For example, the entry in the fourth row, fourth column

[20] Obtaining the true standard error of the difference requires estimating the covariance matrix between the two vectors of estimated parameters. Given the large size of the parameter vector and differences in the sets of the estimated parameters across the models, it was infeasible to estimate the covariance and hence, the true standard error of the coefficient difference.

indicates that when we add three lags of test scores to the model, only about 15 percent of the teacher effect estimates remain the same. Likewise, when prior observable inputs are replaced with three lags of test scores (third row, fourth column), less than 20 percent of the estimated teacher effects remain the same. We see similar results for the models that include a time-constant unobserved effect. Adding two lagged test scores (row nine, column five) results in nearly two-thirds of estimated teacher effects changing significantly. A similar proportion of estimated effects remain constant when three lags of test scores are used in place of lagged observable inputs (row seven, column five).

Second, differences between models that include multiple lags of test scores and those with only a single lagged score are substantial. In the models with no unobserved student/family effect, the three-lags-of-test-scores model (row three, column four) a much lower proportion of teacher effects are statistically indistinguishable (20 percent) from those produced by the baseline model, than the proportion (53 percent) when the three lags of observable inputs are replaced with just the once-lagged test score (row two, column four). Similarly, in models that account for time-constant student/family heterogeneity, we see large differences between the model with three lagged scores and no lagged observable inputs (row seven, column five) and the model with just a single lagged score and no prior inputs (row six, column five).

Third, models in which complete persistence is assumed (and thus all lagged inputs and test scores are removed from the model) produce radically different teacher effect estimates. For models without a time-constant student/family effect, only three percent of the estimates produced by the model with no lagged inputs or scores (first row, fourth column) are statistically indistinguishable from the model with three lags of inputs. Likewise, in the models with a time-constant effect and no lagged scores or inputs (row five, column five), only four percent of teacher-effect estimates are indistinguishable from those generated by the model with three lags of prior inputs.

Finally, among the models with a time-constant effect, first-differenced and fixed-effects models produce rather different estimates of teacher effects. Thirty-seven percent of the teacher effect estimates from the model employing first-differencing and multiple lagged scores (row nine, column five) are similar to those from the three-lagged-inputs model whereas with fixed effects (row eight, column five) only five percent are similar.

G. Correlation of Teacher Rankings across Models

The tests for similarity of coefficient estimates are useful in determining whether the marginal effects of specific teacher credentials vary across model specifications. However, they are less informative for judging the robustness of teacher fixed effects. For individual teacher effects, one is generally less concerned about the specific value of the point estimate. Rather, the relative ranking of teachers is of greater interest, particularly in the context of performance pay systems for teachers. We therefore compare the rank correlation of teacher effect estimates across models in Table 7.[21]

For both models with no unobserved effect and those with a time-constant student/family effect, the highest correlation is between the partial persistence model with one lagged score and the partial persistence model with multiple lagged scores. In both cases the correlations exceed 0.9, suggesting that adding additional lagged scores do not cause a large change in the estimated ranking of teachers.

We also find relatively strong correlations in teacher rankings between models with complete persistence and partial persistence. For models with no unobserved effect (and no lagged inputs), the correlation between the complete and partial persistence models is 0.74. For models with a time constant effect (and no lagged inputs) the correlation in teacher rankings is 0.61. Thus, while

[21] Pairwise correlations are nearly identical to the rank correlations. Estimates of the pairwise correlations are available upon request. Pairwise correlations of the estimated coefficients for the key time-varying covariates are also available upon request.

differences in estimated persistence can have important implications for the efficacy of policies related to teacher quality (Jacob, Lefgren, and Sims 2010), restrictions on the assumed persistence of educational inputs do not result in radically different teacher rankings.

Among the models with no unobserved effect, including lagged inputs in addition to lagged test scores has very little effect on teacher rankings. The correlation in rankings between the model with three lagged scores and no lagged inputs and the model with three lagged scores and three lags of inputs is 0.87. Thus while prior inputs add information, including them in value-added models that exclude a student/family effect does not appear to substantially alter estimated teacher effects. Estimates from models with a time-constant unobserved student/family effect appear to be much more sensitive to exclusion of prior inputs. The correlation in teacher rankings between a model with two lagged test scores and no lagged inputs and one that includes both two lagged test scores and three lags of specific inputs is only 0.40. A possible explanation for this result is that differencing reduces variation in the data and hence, decreases the precision of the estimates and lowers the correlation among them.

For both types of models (no unobserved effect and time-constant effect), excluding prior test scores has a substantial impact on teacher rankings. Among models with no unobserved effect, the most flexible model (including three lagged scores and three lags of inputs) is highly correlated with the model that includes three lagged test scores and no prior inputs (0.87) and the model that includes just one lagged score and no prior inputs (0.83), but has a lower correlation with models that have no lagged scores and three lags of prior inputs (0.62) or no lagged scores and no prior inputs (0.62). The contrasts are, once again, even starker in models that include a time-constant effect. The correlation in teacher rankings between the most flexible model and models with either two or one lagged scores and no lagged inputs are 0.40 or .41, respectively. In contrast, the correlation between the most flexible model and ones with no prior scores and either no or three lags of observable inputs are 0.10 and 0.14, respectively.

The greatest variation in estimates comes when we compare models with differing assumptions about unobserved student/family heterogeneity. For example, the correlation between the model with lagged inputs (but no additional lagged scores) and no control for student/family heterogeneity (column 2) and the same model with time-constant student/family heterogeneity (row 7) is 0.03. Similarly, the correlation between the no heterogeneity model and the model with a time trend for unobserved student/family heterogeneity is 0.15. The correlation between the model with a time-constant effect and the one with a time trend is 0.20. The correlations are higher (0.39) when we compare models which include both multiple lagged scores and inputs (column 5, row 10).

The fact that correlations of estimates from different specifications within the family of time-constant-effect models tend to be lower than those within the group of no-unobserved-effect models and the finding that correlations of estimated teacher effects between time-constant-effect and non-unobserved-effect models are generally low both highlight an inherent trade-off in value-added modeling. As mentioned earlier, controlling for unobserved student heterogeneity with either fixed effects or first-differencing removes a lot of variation from the data, which in turn introduces noise and reduces the precision of teacher effect estimates. By focusing on within-student variation in performance, these models likely to lead to some coefficient estimates being extremely large or extremely small (with large standard errors). The presence of such outliers may in turn result in low correlation coefficients between the estimates.

V. Conclusion

Past research on value-added modeling has been significantly hampered by data limitations, which, in turn, has forced researchers to estimate mis-specified models. The data we use from Florida avoid these limitations and allow for thorough testing of model assumptions and their impact on estimates. We find that common simplifying assumptions are easily rejected. Moreover, the estimated effects of

specific teachers are can be highly sensitive to these mis-specifications. The results are especially sensitive to the assumptions of whether there is a student/family effect and, if so, what form it takes (time-constant versus trending). In models where a time-constant effect is included, teacher effects are very sensitive to other specification choices, e.g., the number of lagged test scores and measured inputs that are included.

Some prior studies have tested some of these assumptions and our results are generally consistent with them. As with Rothstein (2010), we reject strict exogeneity. Also, like Jacob, Lefgren, and Sims (2010), we find that twice-lagged measured inputs affect contemporaneous achievement so that once-lagged achievement is not a sufficient statistic for past inputs. However, we go further in providing evidence that the rate of decay is neither geometric nor equal across inputs. We also reject the assumption that the effects of prior inputs are grade-invariant, although we do find evidence supporting the common assumption that the student/family individual input effect is time-constant.

If the objective is obtaining estimates of teacher productivity with minimum bias, then our evidence suggests avoiding common simplifying assumptions and estimating models that involve more flexible specifications. We find wide variation in estimated teacher effects between more flexible and more restrictive models, including cases with correlations close to zero. Of course, the efficiency of the estimates is also important and imposing these assumptions no doubt reduces the variance of the estimates.

While these results do not provide a great deal of support for the use of teacher value-added measures for high-stakes decisions, they do not necessarily invalidate the idea either. Because other common measures are widely believed to be both more costly and weakly correlated with productivity, a reasonable case can be made for using value-added in some fashion if it provides some signal about true productivity. The lone experimental validation of value-added measures fails to reject the equivalence of differences in value-added estimates of teacher productivity with cross-teacher

differences under random assignment (Kane and Staiger (2008)). Also, several studies now find that teacher value-added measures are positively correlated with the confidential assessments of teachers made by school principals who have arguably the most information about teachers' actual behavior and practice (Harris and Sass 2009; Jacob and Lefgren 2010; Rockoff, Kane, and Staiger 2010). Further, recent simulation evidence suggests that value-added models can produce estimates of teacher productivity that are close to true values under a number of plausible sorting mechanisms (Guarino, Reckase, and Wooldridge 2010).

Given the problems with the assumptions underlying commonly estimated value-added models, caution in using value-added measures of teacher productivity for high-stakes decisions would be advised, but it is still conceivable (and indeed the Rockoff et al. study already suggests) that there may be ways to use the information in conjunction with other measures to increase teacher productivity.

The implications of these results go beyond K-12 education. With the expansion of data collection and data systems, value-added-based accountability analyses are also be considered in higher education (e.g., U.S. Department of Education 2006) and in the health sector (e.g., Werner and Ashe 2005). As computational capacity and data collection capabilities continue to improve, so too will the potential to employ panel data to construct individual performance measures for use in incentive mechanisms. The validity of the assumptions required to create these measures is therefore important to establish and recognize.

References

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

Andrabi, Tahir, Jishnu Das, Asim Ijaz Khwaja, and Tristan Zajonc. 2009. "Here Today, Gone Tomorrow? Examining the Extent and Implications of Low Persistence in Child Learning." Mimeo.

Boardman, Anthony E., and Richard J. Murnane. 1979. "Using Panel Data to Improve Estimates of the Determinants of Educational Achievement." *Sociology of Education* 52(1): 113–21.

Bonesronning, Hans. 2004. "The Determinants of Parental Effort in Education Production: Do Parents Respond to Changes in Class Size?" *Economics of Education Review* 23(1): 1–9.

Chamberlain, Gary. 1984. "Panel Data." In *Handbook of Econometrics*, Vol. II, Z. Griliches and M. D. Intriligator, eds. Amsterdam: Elsevier North-Holland.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2006. "Teacher-Student Matching and the Assessment of Teacher Effectiveness." *The Journal of Human Resources* 41(4): 778–820.

Dee, Thomas S. 2004. "Teachers, Race and Student Achievement in a Randomized Experiment." *Review of Economics and Statistics* 86(1): 195–210.

Ding, Weili, and Steven F. Lehrer. 2007. "Accounting for Unobserved Ability Heterogeneity Within Education Production Functions." Mimeo.

Figlio, David N. 1999. "Functional Form and the Estimated Effects of School Resources." *Economics of Education Review* 18(2): 241–52.

Goldhaber, Dan, and Emily Anthony. 2007. "Can Teacher Quality be Effectively Assessed? National Board Certification as a Signal of Effective Teaching." *Review of Economics and Statistics* 89(1): 134–50.

Guarino, Cassandra M., Mark D. Reckase, and Jeffrey Wooldridge. 2010. "Evaluating Value-Added Models for Estimating Teacher Effects." Mimeo.

Harris, Douglas N. 2007. "Diminishing Marginal Returns and the Production of Education: An International Analysis." *Education Economics* 15(1):31–45.

———. (forthcoming). *Value-Added Measures in Education*. Cambridge, MA: Harvard Education Press.

Harris, Douglas N., and Tim R. Sass. 2009. "What Makes for a Good Teacher and Who Can Tell?" CALDER Working Paper 30. Washington, D.C.: The Urban Institute.

Houtenville, Andrew J., and Karen S. Conway. 2008. "Parental Effort, School Resources and Student Achievement." *Journal of Human Resources* 43(2): 437–53.

Jacob, Brian A., and Lars Lefgren. 2007. "Principals as Agents: Subjective Performance Assessment in Education." *Journal of Labor Economics* 26(1):101–36.

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

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

Koedel, Cory, and Julian Betts. 2009. "Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique." Mimeo.

Lockwood, J. R., and Daniel F. McCaffrey. 2007. "Controlling for Individual Heterogeneity in Longitudinal Models, with Applications to Student Achievement." *Electronic Journal of Statistics* 1: 223–52.

———. 2009. "Exploring Student-Teacher Interactions in Longitudinal Achievement Data." *Education Finance and Policy* 4(4): 439–67.

McCaffrey, Daniel F., J.R. Lockwood, Kata Mihaly, and Tim R. Sass. 2010. "A Review of Stata Routines for Fixed Effects Estimation in Normal Linear Models." Mimeo.

Nye, Barbara, Spyros Konstantopoulos, and Larry V. Hedges. 2004. "How Large are Teacher Effects?" *Educational Evaluation and Policy Analysis* 26(3):237–57.

Rice, Jennifer King. 2003. *Teacher Quality Understanding the Effectiveness of Teacher Attributes*. Washington, D.C.: Economic Policy Institute.

Rockoff, Jonah E., Douglas O. Staiger, Thomas J. Kane, and Eric S. Taylor. 2010. "Information and Employees Evaluation: Evidence from a Randomized Intervention in Public Schools." NBER Working Paper 16240. Cambridge, MA: National Bureau of Economic Research, Inc.

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

Sass, Tim R., Jane Hannaway, Zeyu Xu, David N. Figlio, and Li Feng. 2010. "Value Added of Teachers in High-Poverty and Lower Poverty Schools." CALDER Working Paper 52. Washington, D.C.: The Urban Institute.

Todd, Petra E., and Kenneth I. Wolpin. 2003. "On the Specification and Estimation of the Production Function for Cognitive Achievement." *The Economic Journal*, 113 (485):F3–F33.

———. 2007. "The Production of Cognitive Achievement in Children: Home, School and Racial Test Score Gaps." *Journal of Human Capital* 1(1):91–136.

U.S. Department of Education. 2006. *A Test of Leadership: Charting the Future of U.S. Higher Education*. A Report of the Commission Appointed by Secretary of Education Margaret Spellings, Washington, D.C.

Wayne, Andrew J., and Peter Youngs. 2003. "Teacher Characteristics and Student Achievement Gains: A Review." *Review of Educational Research* 73(1):89–122.

Werner, Rachel M., and David A. Asch. 2005. "The Untended Consequences of Publicly Reporting Quality Information." *Journal of the American Medical Association* 293(10):1239–44.

Wilson, Suzanne M., and Robert E. Floden. 2003. "Creating Effective Teachers: Concise Answers for Hard Questions. An Addendum to the Report: Teacher Preparation Research: Current Knowledge, Gaps, and Recommendations." Washington, D.C.: AACTE Publications.

Wilson, Suzanne M., Robert E. Floden, and Joan Ferrini-Mundy. 2001. *Teacher Preparation Research: Current Knowledge, Gaps, and Recommendations.* Seattle, WA: Center for the Study of Teaching and Policy, University of Washington.

Wooldridge, Jeffrey M. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Wright, S. Paul, Sandra P. Horn, and William L. Sanders. 1997. "Teacher and Classroom Context Effects on Student Achievement: Implications for Teacher Evaluation." *Journal of Personnel Evaluation in Education* 11(1): 57–67.

Tables

Table 1. Summary Statistics for Florida Public School Students in 25 Randomly Selected Districts, 1999/2000-2004/2005

	Mean	Std. Dev.
Student Characteristics		
Female	0.526	0.499
White	0.610	0.488
Black	0.213	0.410
Hispanic	0.129	0.336
Asian	0.026	0.160
American Indian	0.003	0.053
Math Score	692.273	39.289
Math Gain	12.762	23.834
Free/Reduced-Price Lunch	0.373	0.484
Number of Schools	1.028	0.172
Structural Mover	0.191	0.393
Non-Structural Mover	0.134	0.341
Disciplinary Incidents	0.583	1.765
Speech/Language Disability	0.013	0.113
Learning Disability	0.055	0.227
Gifted	0.062	0.241
Mental Disability	0.000	0.219
Physical Disability	0.001	0.035
Emotional Disability	0.003	0.052
Other Disability	0.002	0.048
Limited English Proficiency	0.016	0.126
Teacher Characteristics		
Professional Certificate	0.867	0.340
Ever NBPTS Certified	0.038	0.192
Advanced Degree	0.298	0.457
Years of Experience	10.648	9.524
1 to 2 Years of Experience	0.167	0.373
3 to 4 Years of Experience	0.134	0.341
5 to 9 Years of Experience	0.197	0.397
10 to 14 Years of Experience	0.132	0.338
15 to 24 Years of Experience	0.183	0.387
25 Years Plus	0.120	0.325
Number of Observations	196,015	

Table 2. Tests of Decay Assumptions (Immediate Decay, No Decay, or Geometric Decay) for Models With and Without Student/Family Fixed Effects

	Effect of Student/Family Inputs Decay at Same Rate as Other Inputs			Effect of Student/Family Inputs are Time Constant		
	Immediate (levels - OLS)	None (gains - OLS)	Geometric (levels - OLS)	Immediate (levels - FE)	None (gains - FE)	Geometric (FD - IV)
Lagged Teacher	F(3110,128667) = 2.90 (0.000)	F(3110,128667) = 1.87 (0.000)	F(3110,128667) = 1.82 (0.000)	F(2935,128671) = 3.22 (0.000)	F(2935,128671) = 2.86 (0.000)	F(1989,63173) = 112761 (0.000)
Once Lagged Covariates	F(35,128667) = 100.20 (0.000)	F(35,128667) = 3.72 (0.000)	F(35,128667) = 20.68 (0.000)	F(35,128671) = 3.04 (0.000)	F(35,128671) = 4.14 (0.000)	F(35,63173) = 4.45 (0.000)
Twice Lagged Covariates	F(35,128667) = 25.22 (0.000)	F(35,128667) = 3.83 (0.000)	F(35,128667) = 5.07 (0.000)	F(35,128671) = 2.67 (0.000)	F(35,128671) = 3.11 (0.000)	F(35,63173) = 3.43 (0.000)
Three Times Lagged Covariates	F(35,128667) = 10.92 (0.000)	F(35,128667) = 3.36 (0.000)	F(35,128667) = 5.25 (0.000)	F(35,128671) = 1.90 (0.001)	F(35,128671) = 1.40 (0.059)	F(35,63173) = 1.96 (0.001)
Rate of Decay	1.00	0.00	0.35 (0.000)	1.00	0.00	0.80 (0.094)
Corr (res_t, res_{t-1})	0.504 [0.003]	-0.454 [0.003])	-0.209 [0.004]			-0.505 [0.007]
Strength of the Instruments						F(1,117096) = 29232.13 (0.000)

Notes: Top rows of table report results of *F*-tests where the null hypothesis is that the effects of the inputs reported in the rows are (jointly) equal to zero. Grade, year and school dummies, as well as current period inputs and current period teacher dummies are included in all regressions. Lagged test scores are included in the models with geometric decay. Estimation method is summarized in parentheses under each heading. For example, in the first column, the dependent variable is in achievement levels, and the equation is estimated by OLS. In the last column (FD-IV), the whole equation was differenced and estimated by the instrumental variables estimator with twice-lagged achievement score as the instrument for the differenced first lag of the test score.P-values are reported in parentheses under the test statistics for testing the joint significance of the corresponding variables. The second to last row reports the serial correlation in residuals (standard errors are reported in brackets underneath).

Table 3. Tests for Input Specific Geometric Decay and the Time Constant Unobserved Effect

	Estimation Method			
	First Differencing	Fixed Effects	Trending (FE on FD)	OLS
Reduced/Free Lunch	0.05	0.02	0.12	1.37
	(0.83)	(0.89)	(0.72)	(0.24)
Math Class Size	0.14	2.86	1.33	0.93
	(0.71)	(0.09)	(0.25)	(0.34)
Non-Structural Mover	0.45	0.11	0.00	2.06
	(0.50)	(0.74)	(0.96)	(0.15)
1 to 2 Years Experience	0.48	0.05	0.03	0.14
	(0.49)	(0.82)	(0.86)	(0.71)
3 to 4 Years Experience	0.54	0.71	0.82	0.36
	(0.46)	(0.40)	(0.37)	(0.55)
Advanced Degree	0.00	0.31	0.299	0.71
	(0.99)	(0.58)	(0.08)	(0.40)
Professional Certificate	0.46	0.01	0.97	1.59
	(0.50)	(0.94)	(0.33)	(0.21)
Ever NBPTS Certified	0.30	0.01	0.23	2.71
	(0.59)	(0.94)	(0.63)	(0.10)
Corr ($residual_t$, $residual_{t-1}$)	-0.49			0.51
	[0.08]			[0.03]

Note: Top rows of the table displays the F-statistics and t-statistics for testing the input specific geometric decay. The last row reports the correlation coefficient between the current and lagged residuals. All tests are performed after estimating the baseline model (equation (4)) by the method stated in the heading of the corresponding column. All regressions include grade, year and school dummies, teacher indicators for the current and last periods, as well as three lags of time-varying inputs. P-values are reported in parentheses under the test statistics. Standard errors are reported in brackets under the correlation coefficients.

Table 4. Tests for Age Invariance

	Assumption Regarding Student/Family Inputs (Estimation Method)			
	Decay at Same Rate as Other Inputs (OLS)	Time Constant (FE)	Time Constant (FD)	Trending (FE on FD)
Grade 6 and 7 Once Lagged Covariates	F(70,128671) = 14.40 (0.00)	F(69,128671) = 3.56 (0.00)	F(70,55558) = 3.80 (0.00)	F(36,55558) = 5.87 (0.00)
Grade 6 and Once Lagged Covariates	F(35,128671) = 11.80 (0.00)	F(35,128671) = 2.13 (0.00)	F(35,55558) = 1.97 (0.01)	F(3,55558) = 7.00 (0.00)
Grade 7 and Once Lagged Covariates	F(35,128671) = 10.65 (0.00)	F(34,128671) = 5.33 (0.00)	F(35,55558) = 5.59 (0.00)	F(33,55558) = 5.91 (0.00)
Grade 6 and 7 and Twice Lagged Covariates	F(70,128671) = 6.21 (0.00)	F(70,128671) = 1.60 (0.00)	F(69,55558) = 2.11 (0.00)	F(32,55558) = 4.71 (0.00)
Grade 6 and Twice Lagged Covariates	F(35,128671) = 7.59 (0.00)	F(35,128671) = 2.18 (0.00)	F(34,55558) = 2.97 (0.00)	F(3,55558) = 4.60 (0.003)
Grade 7 and Twice Lagged Covariates	F(35,128671) = 6.78 (0.00)	F(35,128671) = 1.67 (0.01)	F(35,55558) = 2.26 (0.00)	F(29,55558) = 4.48 (0.01)
Grade 7 and Three Times Lagged Covariates	F(35,128671) = 0.93 (0.59)	F(35,128671) = 0.93 (0.59)	F(35,55558) = 1.09 (0.33)	F(35,55558) = 2.27 (0.59)

Note: The table displays the F-statistics for testing age invariance. All tests are performed after estimating the baseline model (equation (4)) augmented by the interactions of inputs with grade dummies. All regressions include grade, year and school dummies, as well as teacher indicators for the current and last periods, as well as three lags of time-varying inputs. The assumptions about the unobserved student/family inputs are stated in the heading of each column, while the estimation method is summarized in parentheses. P-values are reported in parentheses under the test statistics.

Table 5. Tests of Strict Exogeneity

Assumption Regarding Student/Family Inputs (Estimation Method)	Inclusion of Prior Test Scores and Persistence Assumption (Dependent Variable)		
	No Prior-Year Test Score (Achievement level)	Prior-Year Test Score with Complete Persistence (Achievement Gain)	Prior-Year Score with Partial Persistence (Achievement Level)
Unobserved Effect Decays at Same Rate as Other Inputs (OLS)	F(1919,121046) = 5.107 (0.000)	F(1919,94023) = 1.695 (0.000)	F(1919,94023) = 2.555 (0.000)
Unobserved Effect is Time Constant (FE)	F(1534,94025) = 4.398 (0.000)	F(1534,94025) = 3.444 (0.000)	
Unobserved Effect is Time Constant (FD)	F(1530,38231) = 13.382 (0.000)		F(1919,94025) = 1.668 (0.000)
Unobserved Effect is Trending (FE on FD)	F(1577,28229) = 1.117 (0.001)		

Note: All models include: (i) grade, year and school indicators, (ii) current-year time-varying non-teacher inputs and teacher indicators, (iii) prior-year teacher indicators, (iv) three prior years of non-teacher inputs and (v) future teacher indicators. In the regressions with partial persistence, the second lag of the test score is used as an instrument for the differenced first lag of the test score. In the model where the unobserved effect is trending the third lag of non-teacher inputs was not first differenced due to a lack of four-lagged data. P-values are reported in parentheses under the test statistics for testing the joint significance of the future teacher indicators.

Table 6. Test of Similarity

Persistence/ Estimation Method/ No. of Test Score Lags/ No. of Input Lags	Similarity of Coefficients on Nine Time-Varying Characteristics Compared with Estimates from Model with 3 Lags of Inputs and No Prior Test Scores			Similarity of Estimated Teacher Effects (1951 Teachers) Compared with Estimates from Model with 3 Lags of Inputs and No Prior Test Scores		
	No Unobserved Effect	Time Constant Unobserved Effect	Unobserved Effect is Trending	No Unobserved Effect	Time Constant Unobserved Effect	Unobserved Effect is Trending
No Unobserved Effect						
Complete/OLS/0/0	55.6%	77.8%	22.2%	3.0%	30.7%	35.6%
Partial/OLS/1/0	77.8%	77.8%	33.3%	52.9%	41.9%	42.5%
Partial/OLS/3/0	55.6%	88.9%	44.4%	19.5%	47.6%	43.8%
Partial/OLS/3/3	66.7%	88.9%	33.3%	15.5%	39.1%	34.3%
Time Constant Effect						
Complete/FE/0/0	33.3%	0.0%	77.8%	3.3%	3.5%	10.4%
Partial/FD/1/0	66.7%	77.8%	22.2%	0.1%	0.0%	0.1%
Partial/FD/3/0	77.8%	88.9%	11.1%	8.3%	37.3%	75.8%
Partial/FD/2/3	88.9%	66.7%	33.3%	22.0%	37.1%	54.3%

Note: The table reports the percentage of coefficients that are "similar" in the sense that there is no statistically significant difference in the coefficients across specifications. The nine time-varying characteristics" in the left-hand columns are: free/reduced-price lunch status, disciplinary incidents, class size, student mobility, teacher experience levels, teacher educational attainment, and teacher certifications. All models are estimated over a common sample.

Table 7. Rank Correlation of Teacher Estimates Between Models

Persistence/ Estimation Method/No. of Test Score Lags/ No. of Input Lags	Assumption Regarding Student/Family Inputs and Model Specification (Persistence/Estimation Method/No. of Test Score Lags/No. of Input Lags)										
	No Unobserved Effect					*Time Constant Effect*					*Trending*
	C/OLS/0/0	P/OLS/0/3	P/OLS/1/0	P/OLS/3/0	P/OLS/3/3	C/FE/0/0	P/FE/0/3	P/FD/1/0	P/FD/2/0	P/FD/2/3	P/FD-FE/0/3
No Unobserved Effect											
Complete/OLS/0/0	1										
Partial/OLS/0/3	0.0555	1									
Partial/OLS/1/0	0.7422	0.5784	1								
Partial/OLS/3/0	0.7439	0.4996	0.9401	1							
Partial/OLS/3/3	0.6232	0.6229	0.8271	0.8670	1						
Time Constant Effect											
Complete/FE/0/0	0.2738	0.0832	0.2806	0.3056	0.2601	1					
Partial/FE/0/3	0.1583	0.0286	0.1234	0.1238	0.1335	0.6052	1				
Partial/FD/1/0	0.3632	0.1914	0.3340	0.3310	0.4552	0.1471	0.1144	1			
Partial/FD/2/0	0.3462	0.0821	0.2651	0.2373	0.3331	0.1561	0.0750	0.9274	1		
Partial/FD/2/3	0.7028	-0.0200	0.3542	0.3210	0.3906	0.1020	0.1364	0.4069	0.3955	1	
Trending											
Partial/FD-FE/0/3	0.0757	0.1482	0.1505	0.1457	0.1668	0.1473	0.1929	0.2601	0.2534	-0.0103	1


CALDER