EdWorkingPaper No. 22-621

Rethinking Principal Effects on Student Outcomes

Brendan Bartanen
University of Virginia

Aliza N. Husain
Pivot Learning

David D. Liebowitz
University of Oregon

School principals are viewed as critical mechanisms by which to improve student outcomes, but there remain important methodological questions about how to measure principals' effects. We propose a framework for measuring principals' contributions to student outcomes and apply it empirically using data from Tennessee, New York City, and Oregon. We find that using contemporaneous student outcomes to assess principal performance is flawed. Value-added models misattribute to principals changes in student performance caused by factors that principals minimally control. Further, little to none of the variation in average student test scores or attendance is explained by persistent effectiveness differences between principals.

VERSION: August 2022

Rethinking Principal Effects on Student Outcomes

Brendan Bartanen
University of Virginia

Aliza N. Husain
Pivot Learning

David D. Liebowitz
University of Oregon

June 2022

Abstract

School principals are viewed as critical mechanisms by which to improve student outcomes, but there remain important methodological questions about how to measure principals' effects. We propose a framework for measuring principals' contributions to student outcomes and apply it empirically using data from Tennessee, New York City, and Oregon. We find that using contemporaneous student outcomes to assess principal performance is flawed. Value-added models misattribute to principals changes in student performance caused by factors that principals minimally control. Further, little to none of the variation in average student test scores or attendance is explained by persistent effectiveness differences between principals.

***JEL* codes**: I21, J24, J45

Author Note: We thank the following individuals for their helpful comments on the paper: Dale Ballou, Dan Goldhaber, Jason Grissom, Dick Murnane, John Willett, and Jim Wyckoff. We also thank the Tennessee Department of Education, New York City Department of Education, and Oregon Department of Education for providing data access. Lorna Porter provided helpful research assistance with the Oregon data.

1 Introduction

It is widely believed that principals are integral to school performance. Several waves of recent policy reforms—including site-based management, external accountability measures, and teacher evaluation systems—are based on the belief that principals can improve school climate, instructional practices, and student outcomes. With the increasing availability of large-scale longitudinal datasets, a growing literature has used value-added (VA) methods to quantify the impact of effective leadership on student outcomes.[1] These studies consistently conclude that principals' effects are substantial in magnitude. More specifically, they show that variation in school performance (most often conceptualized as students' performance on end-of-year standardized tests) is correlated with who the principal is in a school at a given time, and they interpret this correlation as evidence that higher-quality principals increase school performance.

The core logic of principal VA models is straightforward: by statistically adjusting for factors that affect school performance but that are outside of principals' control, any remaining unexplained variation can be attributed to principal effectiveness. The distribution formed by individual principal VA estimates then provides an indication of the magnitude of "principal effects"—conceptualized as the difference between school performance under the current principal compared to school performance in another plausible setting, such as under a principal of average effectiveness.[2] In practice, however, substantial methodological difficulties exist in credibly identifying the causal effects of school leaders on student outcomes. While prior studies have raised these issues, questions of whether VA models can produce useful measures of principal effectiveness or performance remain unresolved. Can we identify

[1] Examples include Branch, Hanushek, and Rivkin (2012), Coelli and Green (2012), Dhuey and Smith (2014), Grissom, Kalogrides, and Loeb (2015), Laing et al. (2016), and Bartanen (2020). Grissom, Egalite, and Lindsay (2021) survey the literature and find six U.S.-based studies that use panel data to estimate principal value-added measures. The unweighted average of the magnitude of principal effects across the six studies is 0.13 *SD* in math and 0.09 *SD* in reading.

[2] This definition makes explicit that the "importance" of principals is conceptualized as the extent to which variation in the distribution of principal value-added causes higher student outcomes. A different possible conceptualization of importance is to compare student outcomes under a given principal to a counterfactual where there was no principal. Our study does not speak to this latter definition of importance.

high-performing principals using the outcomes of students and schools? How important is variation in principal effectiveness for student learning?

This paper provides new answers to these questions. Specifically, we propose a framework for understanding the contributions of principals to school performance, which we then apply empirically to panel datasets from three distinct contexts: Tennessee, New York City, and Oregon. Collectively, they cover roughly 5 million unique students served by 10,000 unique principals. Our empirical analysis takes two parts. First, we use a variance decomposition approach to establish descriptively how much of the variation in school performance—as measured by student achievement and attendance—is explained by differences between principals. This between-principal variation is the basis for principal VA models, both in terms of understanding the magnitude of principals' contributions to student outcomes and producing estimates of individual principals' effects. The second step of our analysis tests whether the variation attributed to principals by VA models accurately reflects their causal impact on school performance.

Adapting canonical methods for examining "drift" in teacher VA (Chetty, Friedman, and Rockoff 2014a; Goldhaber and Hansen 2013), our proposed framework compares the temporal stability in school performance within the same principal to stability across principals (within the same school). The logic of this comparison is simple: if differences in principal effectiveness are driving persistent or semi-persistent changes in school performance, we should observe that cross-year correlations within the same principal are higher than correlations across principals. Failure to document higher within-principal correlations will lead us to seek explanations for temporal variation in student outcomes from factors that fall outside principals' control. Importantly, our analytic framework allows us to both understand the validity and reliability of principal VA models (as they are currently implemented) and speak to larger questions about the importance of variation in principal effectiveness for student outcomes.

Consistently across all three datasets, we show that cross-year correlations in school performance within and across principals are virtually identical, suggesting that within-school variation in school performance is largely not a function of principal effectiveness. While we replicate existing findings by documenting substantial between-principal variation in school performance, we show that most or all of this variation reflects semi-persistent fluctuations in school-level factors over which we observe principals exerting limited control. In practical terms, our results demonstrate that estimates from principal VA models do not contain accurate information about principal performance or quality. Further, they demonstrate that variation in principal effectiveness does not translate into substantial differences in student test scores or attendance, on average. Ultimately, our results challenge the fundamental paradigm linking effective school leaders to improved contemporaneous student outcomes.

The core measurement challenge with principal VA is that schools have only a single principal at a time and the typical principal remains in a school for only a few years. Events outside of a principal's control—the entrance of a particularly high-performing cohort of students, for instance—create semi-persistent ebbs and flows in school performance that become erroneously attributed to principal effectiveness. This issue is solved neither by statistical adjustment nor the Empirical Bayes shrinkage approaches employed in prior studies. While positive and negative fluctuations would be expected to even out over time, the typical principal's short tenure means that their "value-added" is largely a function of the luck of transient factors they inherit during their tenure. Once we explicitly account for the dynamic nature of school performance, variance decomposition results show that little to none of the variation in student achievement or attendance is explained by persistent differences between principals. Instead, we find that some of the school-level changes in test scores likely reflect compositional changes in teachers and students that would have occurred regardless of which principal was leading the school.

Our results are informative to a broad set of substantive and methodological areas. Understanding principals' contributions—and the magnitude of their contribution compared to that of other educators—provides insights to system leaders on how to prioritize human resource investments in budget-constrained contexts. Whereas prior studies have noted potential concerns related to the construction of principal effect estimates (e.g., Branch, Hanushek, and Rivkin 2012; Chiang, Lipscomb, and Gill 2016; Dhuey and Smith 2018; Cullen et al. 2021), they generally contend that principal VA estimates still contain valuable information about principal effectiveness. We contend that they do not. We argue that existing methods to produce estimates of principal effects on test scores or attendance rates should not be used as tools to measure an individual principal's effectiveness or to collectively describe their importance to the educational production function.

Value-added methodologies are critical to answering a host of questions beyond the overall magnitude of principals' effects. While few local and state agencies use principal value-added models to identify effective school leaders, 46 states use student outcomes as factors in principal evaluation (Donaldson et al. 2021). Prior work has proposed using value-added models as mechanisms to understand the effects of various principal skills, characteristics, and behaviors on teacher and student outcomes (Liebowitz and Porter 2019; Grissom, Egalite, and Lindsay 2021). Additionally, even though the literature that explicitly estimates principal value-added is small, there are countless studies that aim to use changes over time in student-test-score performance to measure the performance of interventions or school processes involving school leaders.[3] Researchers should contextualize future results in the highly dynamic nature of school value-added measures.

2 Conceptual Framework For Estimating Principal Value-Added

The aim of estimating the "value added" of workers in firms emerges from a long tradition in the labor and personnel economics literatures (Koedel, Mihaly, and Rockoff 2015). In the

[3] Just a handful of these types of interventions or school processes include: teacher evaluation schemes (e.g., Taylor and Tyler 2012; Steinberg and Sartain 2015), student suspensions (e.g., Bacher-Hicks, Billings, and Deming 2019; Sorensen, Bushway, and Gifford 2021), and school leader development programs (e.g., Gates et al. 2019; Steinberg and Yang 2022).

context of schools, the goal is to isolate the effect of educators' inputs on student learning in a given year; efforts to do so date back to the 1970s (Hanushek 1971; Murnane 1975). To date, most VA studies have focused on teachers, but there is continued interest in extending these methods to other positions, including principals.

There is strong conceptual support for the notion that principals are a critical input to school performance and, ultimately, student outcomes. Principals are the primary managers of schools whose responsibilities include, for instance, establishing a positive climate, conducting classroom observations and providing feedback to teachers, hiring teachers and other staff, and managing budgets (Grissom, Egalite, and Lindsay 2021; Grissom and Loeb 2011; Liebowitz and Porter 2019). There is also a smaller body of large-scale quantitative evidence using VA methods to link principals to student outcomes (Branch, Hanushek, and Rivkin 2012; Coelli and Green 2012; Dhuey and Smith 2014, 2018; Grissom, Kalogrides, and Loeb 2015; Chiang, Lipscomb, and Gill 2016; Laing et al. 2016; Bartanen 2020). While we discuss the important methodological challenges and potential limitations of these studies below, we note that they are consistent in their findings that variation in principal quality is a meaningful driver of differences in student outcomes. The purpose of our study is to more rigorously evaluate whether VA methods actually measure principal effectiveness.

Fundamentally, we conceptualize principal VA as a variance decomposition exercise. While early research conceived of VA estimation as an effort to understand teachers' contributions to a structural model of educational production, present-day researchers generally specify VA models in an *ad-hoc* fashion and view the resulting estimates as potentially informative of teachers' causal effects (Koedel, Mihaly, and Rockoff 2015). Accordingly, we begin with the premise that VA models are initially descriptive endeavors, and that our estimand of interest does not necessarily have a structural parameter interpretation (Rubin, Stuart, and Zanutto 2004). Our aim is to evaluate the extent to which the descriptive quantities produced by principal VA models contain information about principals' causal effects on student outcomes.

As with any VA approach, the key challenge is to avoid attributing to principals factors that are outside of their control. Particularly for principals, this is a formidable challenge because (1) a school has only one principal at a time, (2) principals cannot reasonably control many school-level factors affecting school performance, (3) *a priori* it is not evident the factors over which principal exert full, partial, or no control, and (4) the typical principal remains in a school for fewer than five years.

To make this discussion more concrete, we decompose the performance (Y) of school s with principal p in year t as follows:

$$Y_{st} = \delta_{pt(s,t)} + \mu_{st} + \nu_{st} \tag{1}$$

where δ denotes principal effectiveness, μ denotes other school factors over which principals are theorized to have minimal or no control and should not contribute to estimates of their effectiveness, and ν_{st} is a random error term capturing purely transient factors.[4] As with teacher VA, Y is most often conceptualized as a measure of average student test score performance in year t, with adjustments for baseline factors such as students' demographic characteristics and prior-year test scores.[5] The most important conceptual difference between estimating principal effects and teacher effects is that principals do not provide direct instruction in classrooms, and thus their impact on student achievement (δ) is largely mediated by school-level processes.[6] δ may include, for instance, a principal's efforts to recruit and retain high-quality teachers, or their ability to establish a positive school climate.

As noted above, a major challenge for estimating principal VA is that there are likely other school-level factors over which principals have limited or no control. These are captured by

[4] For parsimony, we largely refer to μ as factors that principals cannot control, though we acknowledge that principals likely have *partial* control over many school-level processes. Thus, a given process (e.g., teacher hiring) might operate both through μ and δ.

[5] To keep the focus on our conceptual argument, we leave considerations of how to construct Y for the methods section.

[6] There are also some direct channels by which principals can affect test-score performance such as via directly motivating students or through a role-model effect. Our models are unable to parse these direct and indirect pathways.

μ in Equation 1. Because a school has only one principal at a time, it is difficult to separate μ from δ. For example, while effective principals may be able to better identify high-quality applicants for open positions, they likely face constraints over the hiring pool, which is a function of uncontrollable factors like geography, local labor market conditions, and the salary schedule. While these factors may be partially captured by standard observables in administrative datasets, such as a school's average student demographics, there likely remains a substantial portion of μ that is unobserved. The typical approach in prior work is to account for μ by estimating principal VA via a model with principal and school fixed effects (e.g., Branch, Hanushek, and Rivkin 2012; Dhuey and Smith 2014, 2018; Grissom, Kalogrides, and Loeb 2015; Laing et al. 2016; Bartanen 2020). In this model, principal fixed effects (i.e., their VA estimates) are identified by comparing principals who worked in the same school in different years. Assuming some principals worked in multiple schools, the model further allows comparisons to be among principals in "connected networks," in which every school has had at least one principal move to at least one other school in the network (see Bartanen and Husain 2021).

The key identification assumption of the principal and school fixed effects approach is that there are no time-varying unobserved school factors that principals cannot control. More explicitly, prior studies assume that $\mu_{st} = \mu_s$, such that all persistent or semi-persistent within-school changes in school performance are attributed to δ. But this is an incredibly strong assumption—schools are complex organizations with students, teachers, staff, and parents interacting with one another and with the broader ecosystem (e.g., the neighborhood). Principals also inherit the conditions set by their predecessor(s), such as a large proportion of the teaching staff. It is likely, then, that μ has both fixed and dynamic components.

To allow for the possibility of these dynamic components, we modify Equation 1 as follows:

$$Y_{st} = \delta^F_{p(s,t)} + \delta^D_{pt(s,t)} + \mu^F_s + \mu^D_{st} + \nu_{st} \tag{2}$$

where δ and μ have both fixed (F) and dynamic (D) components. Note that $\text{cov}(\delta^F, \delta^D) =$

$\text{cov}(\mu^F, \mu^D) = 0$, by construction. Following Chetty, Friedman, and Rockoff (2014a), we assume that both δ^D and μ^D fluctuate stochastically over time according to a stationary process:

$$\begin{aligned} E[\delta^D_{pt(s,t)}|t] = E[\mu^D_{st}|t] = E[\nu_{st}|t] = 0, \\ \text{cov}(\delta^D_{pt(s,t)}, \delta^D_{p,t+x(s,t)}) = \sigma_{\delta^D_x}, \ \ \text{cov}(\mu^D_{st}, \mu^D_{s,t+x}) = \sigma_{\mu^D_x}, \\ \text{cov}(\nu_{st}, \nu_{s,t+x}) = \sigma_{\nu_x}, \ \ \text{for all} \ \ t \end{aligned} \tag{3}$$

where x denotes the difference between years for a given t. The stationarity assumption is non-trivial. It requires that within principal-by-school spells, the stability of school performance between year t and $t + x$ depends only on x (the number of years separating the school-by-year cells). As we describe further below, this assumption allows us to isolate the principal's contribution to school performance by comparing the cross-year stability of school performance within and across principals.

In practical terms, μ^D in Equation 2 allows for semi-persistent fluctuations in school-level factors that principals cannot control. Two examples of such fluctuations are teacher composition and the readiness of incoming cohorts of students. If a highly effective teacher retires for reasons outside the principal's control (e.g., full eligibility for pension benefits), school performance will decline, but this is not a one-time "shock" (which would be fully contained in ν_{st}) because their replacement will likely remain in the school for multiple years. Equally, the performance of a middle school's "feeder" elementary school may change over time, leading to variation in the readiness of incoming cohorts of students. Again, this is not a one-time shock because students remain in the school for multiple years. Unless this variation is fully captured by observable baseline characteristics (e.g., prior test scores), differences in cohorts' unobserved achievement-gain potential will contribute to μ^D.

The presence of μ^D creates a problem for isolating a principal's contribution to school performance because a school has only one principal at a time. This means that in the typical school fixed effects approach, changes in Y conflate μ^D and $\delta^F + \delta^D$. In other words, we do not know whether within-school changes in performance are caused by principals or by time-varying factors that they cannot control. As a principal's tenure length increases, positive and negative fluctuations in school performance caused by μ^D will even out, in expectation. The typical principal, however, remains in a school for just a few years, creating the potential for substantial small sample bias that hinges on the magnitude of μ^D.

Differences in tenure length across principals can shed light on the importance of μ^D. The basic intuition of our approach is to examine the stability of school performance for sub-samples where the school's principal in a given year is the same or different as the principal x years later. The difference in these correlations provides a lower bound estimate on the magnitude of principals' contributions to school performance.[7] To see this, we can write these correlations as:

$$\begin{aligned} r_x^{\text{SamePrin}} &= \frac{\sigma^2_{\mu^F} + \sigma_{\mu^D_x} + \sigma^2_{\delta^F} + \sigma_{\delta^D_x}}{\sigma^2_Y}, \\ r_x^{\text{DiffPrin}} &= \frac{\sigma^2_{\mu^F} + \sigma_{\mu^D_x} + \sigma_{\delta^F_x}}{\sigma^2_Y}, \\ \text{where } \; \sigma_{\delta^F_x} &= \operatorname{cov}(\delta^F_{j(s,t)}, \delta^F_{k(s,t+x)}) < \sigma^2_{\delta^F} \; \text{ for } \; j \neq k \end{aligned} \tag{4}$$

We assume that $\operatorname{cov}(\mu^D_x, \delta^D_x) = \operatorname{cov}(\mu^D_x, \delta^F_x) = \operatorname{cov}(\mu^F_x, \delta^D_x) = 0$, meaning that the dynamic components of δ and μ are uncorrelated with each other and with the respective fixed components. We, however, allow for the possibility of nonrandom sorting of principals to schools on the basis of their fixed components. This is represented by $\sigma_{\delta^F_x}$ in r_x^{DiffPrin}, which is the covariance between the effectiveness of principals j and k for school s. A positive covariance, for instance, will increase the stability of school performance across different principals. Prior evidence on teacher sorting suggests, however, that this will likely be modest. For instance, Chetty, Friedman, and Rockoff (2014a) find that 85 percent of the variation in teacher VA is within rather than between schools.

[7] This framework is similar to that used by Branch, Hanushek, and Rivkin (2012), who regress the squared difference in residualized achievement gains between year t and t^* on an indicator for whether the principal is different in those two years. Whereas they pool across all available pairs of years and do not account for the difference in time, we directly incorporate the potential for drift by producing an estimate for each value of x.

Intuitively, $r_x^{\text{SamePrin}} - r_x^{\text{DiffPrin}} = \frac{\sigma^2_{\delta^F} + \sigma_{\delta^D_x} - \sigma_{\delta^F_x}}{\sigma^2_Y}$ provides insight about the extent to which school performance is driven by principals versus school-level factors that principals cannot control. This comparison is similar to a difference-in-differences logic. That is, we are assuming that the variance of μ^D is the same for these two sub-samples, which follows from the stationarity assumption in Equation 3. This allows for $r_x^{\text{SamePrin}} - r_x^{\text{DiffPrin}}$ to isolate the extent to which principals cause persistent or semi-persistent changes in school performance. If principals are systematically driving these changes, then $\sigma^2_{\delta^F} + \sigma_{\delta^D_x} - \sigma_{\delta^F_x} > 0$ and $r_x^{\text{SamePrin}} > r_x^{\text{DiffPrin}}$. How this difference varies as a function of x is also informative about the stability of principal performance (i.e., the magnitude of δ^F_p relative to δ^D_{pt}). If the dynamic component of principal effectiveness is small, the difference between r_x^{SamePrin} and r_x^{DiffPrin} should be similar for all x. If the dynamic component is large, this difference should be larger when x is smaller because current principal performance is a less reliable predictor of future performance. With this framework established, we turn now to our empirical work.

3 Data, Sample, and Measures

This study analyzes longitudinal administrative data from two mid-sized states and the largest school district in the United States. All three data sets contain detailed information about all employees in the K–12 public school system, including job title, school placement, and demographic information. We connect these staff data to student files which include demographic and enrollment information, as well as achievement scores on statewide end-of-year exams. We provide brief information in the main text on our samples and outcome measures, and refer readers to Appendix B for further details on the data from each context.

3.1 Sample

The Tennessee data, provided by the Tennessee Department of Education via the Tennessee Education Research Alliance at Vanderbilt University, cover the 2006–07 through the 2018–19 school years, and (in their most comprehensive sample) represent 4,095 unique principals, 19,867 school-by-year cells, and 10.0 million student-year observations. The New York City (NYC) data from the New York City Department of Education cover the 1998–99 through the 2016–17 school years, and represent up to 3,201 unique principals, 18,240 school-by-year cells, and 6.2 million student-year observations. The Oregon data, provided by the Oregon Department of Education, cover the 2006–07 through the 2018–19 school years, and represent up to 2,757 unique principals, 12,449 school-by-year cells, and 5.4 million student-year observations. Across all contexts, our sample represents roughly 10,000 principals, 5 million unique students, and 22 million student-year observations. As we discuss below, one important limitation of the NYC data is that we cannot access an underlying student enrollment or attendance file, which means we cannot conduct certain analyses that we show for TN and OR. We report the characteristics of the students and principals in our samples in Appendix Tables B.1–B.3.

3.2 Outcomes

The primary measures we study are school-by-year level means of students' contemporaneous test-score results in math and reading. These test scores are available in grades 3–8 for all contexts, but we also examine high school students' exams in TN and OR.[8] We also examine their daily attendance rates in auxiliary models. Within each dataset, these student outcomes

[8] In Tennessee, end-of-course exams are required for various math and reading courses, including Algebra I and II, and English I, II, and III. Through the 2013–14 school year, Oregon required high-school students to sit for the Oregon Assessment of Knowledge and Skills (OAKS) in math and English Language Arts at some point in high school; students across grades 9–12 sat for the test. The state shifted to the Smarter Balanced Assessment Consortium (SBAC) test in 2014–15 at which point all 11th-graders were required to sit for the test. Thus, students are typically tested one time in high school. See the data appendix (Appendix B) for further information about HS exams.

are standardized at the grade-by-year level to have a mean of zero and standard deviation of one, and we report estimates of the magnitude of principal effects in student-level standard deviation units.[9] We describe the specific construction of school performance measures below.

4 Analytic Approach

Our general approach takes two steps. First, we use variance decomposition to descriptively examine how much of the variation in student outcomes is explained by differences between schools, differences between principals within schools, and differences within principals over time. This first step is important, as obtaining credible measures of principal effectiveness assumes that there exists a distribution of principal quality (with respect to raising student test scores) with nonzero variance. Once we establish the magnitude of these variance components, we then evaluate the credibility of empirical estimates of the principal-level variance component as measures of principal quality, effectiveness, or performance.

4.1 Variance Decomposition

To conduct our variance decomposition, we first obtain student-level test score residuals by regressing student test scores on a vector of observable characteristics:

$$\begin{aligned} Y_{ist} &= \beta \mathbf{X}_{ist} + \gamma_s + \epsilon_{ist} \\ Y^*_{ist} &= Y_{ist} - \hat{\beta}\mathbf{X}_{ist} \end{aligned} \tag{5}$$

We estimate β using within-school variation by including a school fixed effect (γ_s), which avoids overstating the impact of observables on student test scores due to a potential correlation between $\mathbf{X}_{ist}$ and school quality. We then compute $\bar{Y}^*_{st}$, the school-by-year mean of test score residuals (Y^*_{ist}), which is our measure of school performance. We exclude school-by-year

[9] To account for the different grades at which high-school students take these tests, we standardize their scores across grade levels, within the high-school grade band.

cells where fewer than 25 students contribute to $\bar{Y}^*_{st}$.[10] Finally, we estimate a random effects model to partition the variance in school performance into differences between schools, differences between principals nested within schools, and differences between school years nested within principals:

$$\bar{Y}^*_{spt} = \theta_s + \theta_{s,p} + \epsilon_{spt} \tag{6}$$

where θ_s is a school random effect, $\theta_{s,p}$ is a principal-by-school random effect, and ϵ_{spt} is an i.i.d. error term that parses out purely transient factors of yearly school performance, such as test-score measurement error or a fire alarm on the day of the exam. The parameters of interest are their estimated variances. In particular, we are interested in the the variance of $\theta_{s,p}$, which is the magnitude of the variation attributed to principals.[11] Implicitly, Equation 6 follows the logic of prior principal VA studies in attributing all persistent or semi-persistent changes in school performance to principals.

A crucial decision in this approach concerns the appropriate elements of $\mathbf{X}_{ist}$, which are determinants of (or proxies for) student test scores that should not be attributed to school performance. In the teacher value-added literature, $\mathbf{X}_{ist}$ typically includes prior-year test scores and absences, student demographic and academic characteristics (e.g., gender, race/ethnicity, economic disadvantage, special education status, limited English proficiency), class- and school-by-year means of the individual student characteristics, and fixed effects for grade and year. In particular, including prior-year outcomes is important to control for dynamic sorting of students to classrooms and teachers, and including classroom-level

[10] As noted in the data description (Appendix B), this drops only a few thousand students, largely in schools that do not cover tested grades (e.g., K–2 schools) but where a handful of students had recorded test scores.

[11] Note that while some principals work in multiple schools, we are not leveraging this potential source of variation in our primary models because we treat principals as perfectly nested within schools. While estimating Equation 6 using a cross-classified model (i.e., with θ_p instead of $\theta_{s,p}$) could help to disentangle principal-to-school sorting that could lead to inflation of the magnitude of θ_s, it requires fairly strong assumptions about the nature of principal-school complementarities to justify the transitivity of a principal's impact in different schools (Bartanen and Husain 2021). Additionally, there are relatively few principals whom we observe in multiple schools. Since our primary aim is to understand the nature of within-school comparisons of principals, we opt for the nested model. However, we show as extensions of our main results estimates that examine the stability of estimated principal effects across schools.

controls is important to account for peer effects (Rothstein 2010; Chetty, Friedman, and Rockoff 2014a). These are particularly salient given empirical evidence on the phenomenon of parental requests for their children to be assigned to particular teachers (e.g., Jacob and Lefgren 2007). For school or principal value-added, there is no consensus on the appropriate set of controls.

Given the lack of consensus about the appropriate controls and the considerations we outline in the next paragraph, we examine five specifications in our preliminary step of residualizing student test scores, prior to using these residualized values to decompose the remaining observed variance. Model 0 includes no controls. Model 1 includes observable student characteristics, school-by-year averages of these characteristics, and fixed effects for grade and year.[12] Model 2 adds cubic polynomials for students' prior-year test scores in math and reading, as well as a cubic for their prior-year attendance rate. Model 3 repeats this specification but restricts the sample to students who are in their first year in the school. Model 4 replaces prior-year outcomes with prior-school outcomes, which is defined as the most recent prior-year outcome where the student was in a different school.[13] In each case, the residualization is performed in a model with school fixed effects to avoid over-controlling for true differences in school quality that may be correlated with elements of $\mathbf{X}_{ist}$, such as student demographics.

Each of these models has differing strengths and limitations that encompass both conceptual and practical considerations. We examine Model 0 mainly for the sake of comparison to demonstrate the relative importance of controlling for the elements in $\mathbf{X}_{ist}$. Model 1 accounts for school-level sorting on the basis of observable student and family characteristics. By omitting prior test scores, Model 1 avoids the problem of controlling away part of

[12] In Tennessee and New York City, these student characteristics include gender, race/ethnicity, parental income (as measured by eligibility for free- or reduced-price lunch), special education status, and English learner status. Oregon additionally includes 504 plan designation and participation in migrant or Indian education programs.

[13] We do not estimate Models 3 and 4 for NYC because, while we can observe students in schools where they take their math and reading tests, we do not have the requisite enrollment information that allow us to observe when students first enter a school or when students move in and out of schools.

a school's repeated effect on student performance, but it will potentially under-control for sorting. As long as any student-to-school sorting on unobservables is fixed over time, however, the bias will be limited to the school-level variance component. Similarly, principal VA models that include school fixed effects will control for any time-invariant student-to-school sorting (whether based on observables or unobservables).

By including prior-year outcomes, Model 2 is the most aggressive approach in terms of accounting for the myriad potential factors that affect student outcomes in year t but that should not be attributed to school or principal performance. Effectively, prior-year outcomes are intended to serve both as a sufficient statistic for each student's history of inputs (in- or out-of-school) up to year $t - 1$ and a proxy for unobserved student characteristics, such as motivation. The disadvantage of Model 2, however, is that it will control away part of a school's causal impact on student performance. This adjustment will lead to a downward bias in the principal- and school-variance components and will punish high-performing schools (or reward low-performing schools). Nevertheless, the substantive importance of this bias is not immediately clear and may be outweighed by the benefit of more aggressively adjusting for sorting. Among these approaches, Model 2 is most closely aligned with teacher VA and is also the most common approach in the principal VA literature.

The final two models aim to find middle ground. By controlling for prior-year outcomes but only including new-to-school students, Model 3 avoids the repeated effects issue. The obvious cost of this approach is that it greatly reduces the sample size, which may lower the reliability of VA estimates and may introduce external validity concerns if school or principal quality matters differentially for new-to-school students. Model 4 replaces prior-year outcomes with each student's most recent prior outcome that was in a different school. This is conceptually similar to Model 3 but has the benefit of a larger sample size, though it still fails to include most elementary school students.

4.2 Validity and Reliability Analyses

After establishing variance components for school performance, we then investigate the extent to which the within-school variation attributed to principals ($\theta_{s,p}$) is a valid and reliable measure of principal effectiveness. Put more simply, do within-school changes over time (net of purely transient fluctuations) in mean student test score residuals reflect the causal effect of principals or of factors outside their control?

Following the framework of Equation 4, we estimate the correlations among pairs of school-by-year mean test score residuals ($\bar{Y}^*_{spt}$) constructed from Equation 5, precision weighting by the total number of students in each pair of school-by-year cells. We use all available school-by-year cells with a time span of x years between them. These autocorrelations, r_x, represent the reliability (also described as "stability") of mean school-by-year test scores for predicting school performance x years later (Chetty, Friedman, and Rockoff 2014a). In the context of teachers, prior work finds that the correlations decay as x increases up to roughly seven years, but are stable afterwards, implying that teacher quality has permanent, dynamic, and transitory components (Goldhaber and Hansen 2013; Chetty, Friedman, and Rockoff 2014a). We estimate r_x^{SamePrin} and r_x^{DiffPrin}, which are correlations across school-by-year mean test score residuals where the principal in year t is the same or different as year $t + x$. Their difference indicates whether principals contribute to changes in school performance.

5 Results

We begin by establishing variance components for school performance across contexts and approaches to residualization, which illustrates the basis for prior claims that principals matter for student outcomes. We then move to the heart of our analysis, which evaluates whether these descriptive quantities accurately reflect principals' causal effects as opposed to factors outside of their control.

5.1 Variance Decomposition

In Table 1, we decompose the total variance in school-by-year mean achievement into three components: between-school, between-principal (nested within school), and within-principal (across years). We show these decompositions across our five models for residualizing student test scores and across our three datasets. For each level (school, principal, residual), we report the estimated standard deviation of the random effect, with the variance component (%) shown below. For parsimony, we focus on the results for math scores, with results for reading (which are very similar) shown in Appendix Table A.1.

These results demonstrate that roughly 10 to 20 percent of the observed variation in school math performance is attributed to principals, with some variation across contexts and residualization specifications. It is notable, however, that the residual variance component is substantial across all models—in each case it is larger than the principal variance component. In Model 2, which aims to measure student achievement "growth" by including adjustments for prior-year test scores, the residual dwarfs both the school and principal variance components, demonstrating that "school effectiveness" (as measured by student test score improvement) varies quite substantially across years, even within the same school and principal.

Comparing the results across models is informative to understand the potential strengths and weaknesses of different approaches for residualizing student test scores. Comparing Model 0 (no controls) and Model 1 (controls for student demographics), we observe a reduction in the magnitude of each of the variance components, but by far the largest change is for the school component. By contrast, the change in the principal variance component is substantially smaller (NYC) or roughly zero (TN and OR). This shows that student sorting—as measured by student demographic characteristics—is largely between schools, as opposed to students or families responding to principal changes within schools.

Consistent with our expectations, we find that controlling for prior-year test scores in Model 2 further reduces the magnitude of the school and principal effects. As previously discussed, part of this reduction is due to mechanically controlling for the school or for the principal's own quality for students who remain in the same school across years. The reduction may also, however, encompass further elimination of non-random student sorting that was not captured by student demographics. For instance, a middle school's "feeder" elementary school might be particularly effective, which increases the readiness of incoming student cohorts. This increased readiness may be orthogonal to student demographics and will lead to higher test score performance in the middle school, but should not be credited to the middle school or its principal. Whether the benefit of controlling for prior test scores outweighs the cost is unclear.

To try to disentangle this issue, we can repeat this prior-year test score specification for a sample of students who are in their first year in the school (Model 3), or instead control for students' most recent prior-year score that was in a different school (Model 4). We find similar results for both of these models. As expected, the school and principal effects increase in magnitude relative to Model 2, though only modestly. In particular, the school random effect remains substantially smaller in magnitude than in Model 1, suggesting that prior test scores are capturing additional between-school-sorting that is not fully accounted for by student demographics. Additionally, the fact that the principal effect remains smaller in Models 3 and 4 than Model 1 suggests that there is also time-varying heterogeneity within schools over time that is not completely captured by controlling for student demographics. This is important to establish because it means that not controlling for prior-year test scores (Model 1) will likely over-attribute changes in test score performance to principals and schools, while controlling for prior-year test scores will tend to understate their effects.

The results for Models 3 and 4 also illustrate the practical costs of these alternative approaches. Whereas we can estimate Models 0–2 on a full and stable sample, we apply Models 3 and 4 to restricted samples. In some cases, Models 3 or 4 are simply intractable due to data limitations (e.g., not observing the full history of student enrollment in NYC) or the fact that very few students in early grades will have test scores from a prior school. We present analogous results for reading (Appendix Table A.1) and (in Tennessee and Oregon) for attendance (Appendix Tables A.2 and A.3). The results are substantively identical, with smaller estimated variances attributed to schools and principals for reading than math.

To summarize, Table 1 establishes two important points. First, regardless of the specification, the non-zero magnitude of the principal random effects demonstrates that there exists within-school variation in school performance that is correlated with principal assignment. This is the source of variation that existing studies leverage to estimate principal value-added, but it is still unclear whether this variation reflects the causal effects of principals on student outcomes. Second, even in models that leverage hundreds of student-by-year observations to estimate each school or principal effect, there remains substantial residual variation in school performance across years. In some of the models that adjust for prior test scores, the magnitude of year-to-year fluctuations in school performance outweighs the stable components of schools and principals.

5.2 Validity and Reliability

These findings motivate the next part of our analysis, which seeks to understand the extent to which within-school variation in school performance across principals is a valid and reliable measure of principal performance. As a first step, we follow prior canonical studies of teacher value-added (Goldhaber and Hansen 2013; Chetty, Friedman, and Rockoff 2014a) in computing the correlation between yearly mean test score residuals ($\bar{Y}^*_{spt}$) within schools. For teachers, these autocorrelations (between teacher-by-year rather than school-by-year cells) indicate both the stable and dynamic nature of teacher VA. Chetty, Friedman, and Rockoff (2014a) and Goldhaber and Hansen (2013) both find adjacent-year correlations well below 1, demonstrating the large role of estimation or measurement error in teacher VA. They also find declining correlations between mean residuals that are further apart in time, suggesting that teacher effectiveness has a substantial dynamic component that "drifts" over time. We conduct the same exercise for schools, but also examine whether the patterns differ for mean residuals within the same principal versus those across principals (within the same school).

Figure 1 shows these within-school autocorrelations across each of the four residualization models, weighting by the total number of students used to form the mean residual in each cell. We show results for math and reading in each context. We uncover several important patterns. First, while the magnitude of the correlations varies across models (reflecting the magnitude of the school-level variance component relative to the residual), they all follow the same pattern of declining correlations as the time span between outcomes grows larger. Particularly in Models 2–4, which control for students' prior test scores, school performance in prior years quickly becomes only weakly predictive of current performance. A modest positive correlation remains, however, even when comparing performance with a 10-year gap. While this pattern of declining correlations is similar to prior findings for teacher VA, it is perhaps more striking given that school VA measures should contain substantially less estimation error due to the large number of students whose test scores contribute to the estimate.

The autocorrelation vectors in Figure 1 demonstrate that school performance has a substantial dynamic component. This is unsurprising from a conceptual perspective, as the school factors that matter most for student achievement are not fixed over time. In particular, the school's personnel—teachers, administrators, and other staff—are changing, as are external factors like neighborhoods and district policy. The key question for our analysis is to what extent these changes in school performance over time are driven by principals or by other factors over which principals exert little control? To shed light on this, we compare the correlations between year t and year t^* on subsamples of years where the principal in year t is either the same or different as year t^*.

We find strikingly small differences in school performance autocorrelations comparing within-principal spells, as compared to across principals. We show these correlations in Figure 2. For parsimony, we focus on the correlations from Model 2 (residualizations that include prior-year test scores), with results for other models shown in Appendix Figures

A.1–A.3.[14] While the correlations within the same principal tend to be slightly larger in magnitude, particularly after the first few lags, the pattern of decreasing correlations in school performance across time is largely *not* explained by principal transitions. Based on $r_1^{\text{SamePrin}} - r_1^{\text{DiffPrin}}$, the estimated *SD* of principal VA in math is 0.035, 0.029, and 0.034 in TN, NYC, and OR, respectively. These estimates are substantially smaller than both those reported in prior studies and based on the variance decomposition in Table 1.[15] Except for math in OR, $r_x^{\text{SamePrin}} - r_x^{\text{DiffPrin}}$ effectively approaches zero as x increases. For reading, there is virtually no difference across all three contexts between same-principal and different-principal correlations for any x.

The results in Figure 2 show that the dynamic component of school performance is not driven by differences in student outcomes across principal tenures. Instead, there exists within-school variation in student achievement performance that is semi-persistent, and, thus, becomes erroneously attributed to "effectiveness" differences between principals. As a result, most, if not all, of the within-school variation in student test scores used to produce principal VA estimates is likely unrelated to principal performance. While there is some suggestive evidence in the math results for OR that the correlations within principals are greater than between principals after the first few lags, it is important to understand that the correlations with higher lags rely on fewer principals, particularly in OR (see Appendix Table A.4). This makes the estimated correlations noisier and also introduces a potential sample selection effect, as the vast majority of principals we examine remain in the same school for fewer than five years. That said, it could also indicate a scenario where new-

[14] We show corresponding autocorrelations vectors for attendance in Tennessee and Oregon in Appendix Figures A.4 and A.5, respectively.

[15] Perhaps most notably, these *SD* estimates are roughly 40–60 percent smaller than the lower-bound estimates from Branch, Hanushek, and Rivkin (2012). We provide a replication of the Branch, Hanushek, and Rivkin (2012) results using our datasets in Appendix Table A.5. Specifically, we construct the squared difference in mean residuals (from the Model 2 approach) between each possible pair of school-by-year cells. We then regress these squared differences on an indicator for whether the principal is different between these two cells. We obtain results comparable to theirs, but also show how the presence of drift yields an inflated estimate of the magnitude of principal effects. Specifically, we add non-parametric controls for the time gap between pairs of school-by-year mean residuals. Because of drift, these indicators are positive and large in magnitude. They are also highly correlated with the different principal indicator and, by consequence, controlling for them greatly attenuates the coefficient on different principal.

to-school principals' effects are small in magnitude and grow over time. We return to this possibility in the next section.

As a useful point of comparison, we show estimates of r_x^{SamePrin} and r_x^{DiffPrin} using perception-based measures of principal performance. In NYC and TN, we can examine for a subset of years rubric-based ratings from their supervisors and low-stakes survey-based ratings from their teachers.[16] We do not argue that these rating scores are intended to measure the same thing as student test score performance or even that they are better measures of principal performance, but they are informative with respect to illustrating the decomposition logic of the correlation analyses. Figure 3 shows within-school autocorrelations of these measures, again comparing within the same principal versus across principals. Here, we observe clear separation in the correlations within versus across principals: correlations within the same principal are substantially larger than correlations across principals. This demonstrates that there are substantial differences in the *perceptions* of principal performance, which may also reflect an important dimension not captured by student-outcome-based measures.

5.3 Checking the Stationarity Assumption

Fundamentally, our proposed framework rests on the idea that principal effects exhibit some persistence—either fixed or evolving through time. Failure to observe higher stability of school performance within versus across principals, then, indicates that variation in principal effectiveness is not driving changes in school performance. Perhaps the most important potential threat to validity under this framework is that the stability of a principal's performance—or, equally, uncontrollable school factors—varies as a function of how long the principal has been leading the school. A common suggestion in the principal VA literature, for example, is that some of a principal's effect is lagged, such that school performance under a new principal is more reflective of the conditions established by her predecessor(s) than

[16] In TN, each of these measures becomes available starting in the 2011–12 school year, meaning that we can examine gaps of up to seven years. In NYC, only four years of survey data are available, such that the maximum gap is three years.

of her own performance. As she remains in the school, however, her influence over school performance increases. Under this scenario, the stationarity assumption would be violated and $r_x^{\text{SamePrin}} - r_x^{\text{DiffPrin}}$ would likely understate the magnitude of principal effects for small values of x. Specifically, the partial persistence of the prior principal's effect would decrease r_x^{SamePrin} and increase r_x^{DiffPrin}.

We examine this empirically by comparing r_x^{SamePrin} and r_x^{DiffPrin} for principals of varying tenure levels. If the prior principal's effect persists into the new principal's tenure, r_x^{SamePrin} should be smaller for principals in their first few years in the school, relative to more established principals. By similar logic, r_x^{DiffPrin} should be larger when the principal in year t (i.e., the departing principal) has a longer tenure. We show these results for each context in Appendix Figures A.6–A.8.[17] We do find some evidence that year-to-year correlations are slightly greater among the longest-tenured principals in Oregon (math and reading) and New York City (math only). The sample sizes for these cells, however, are also quite small, particularly in Oregon where very few principals stay in the same school for more than five years (see Appendix Tables A.6–A.8). In New York City, the larger correlations are only observed for principals with 16 or more years of tenure (and not for 11–15 years), which is a small and potentially idiosyncratic group. In Tennessee, there is little difference across tenure groups and, if anything, the correlations are smaller among highly tenured principals.

Overall, these results suggest that the limited variation in measured effects of principals is not merely a function of a preponderance of new-to-school principals whose potential impacts are smaller than longer-tenured principals. We provide further evidence in support of this claim below.

[17] Note that these figures only report estimates for cells where we observe at least 50 principals. The sample size for each context is shown in Appendix Tables A.6–A.8.

5.4 Event Studies

Given the implicit difference-in-differences logic in comparing the stability of school performance within versus across principals, we extend our analysis in an event study framework. Specifically, we examine changes in school performance when a principal exits from a high-performing or low-performing school, and we compare these observed changes to a set of comparison schools that had similar performance in the pre-period but did not change principals. This allows us to test whether the changes in school performance following a principal transition—variation which is leveraged to estimate a principal's value-added—correctly reflects the causal effect of a principal as opposed to time-varying school factors that would have happened even if the school kept the same principal.

To implement this analysis, we first identify sets of high-performing and low-performing schools using a three-year moving average of school-by-year mean test score residuals ($\bar{Y}^*_{spt}$). We define high- and low-performing schools as the top and bottom quartiles using this measure. Within each of these groups, we find schools that changed principals following the current year and construct a six-year panel defined by event time -3 to 2, where 0 denotes the first year of the new principal and the performance quartile is generated from $\bar{Y}^*_{spt}$ in time -3 to -1.[18] We further restrict that the school must have had a single principal in each pre- and post-period (i.e., the departing principal was in the school for at least three years and the new principal stayed for at least three years). With this sample, we regress $\bar{Y}^*_{spt}$ on a set of indicator variables for event time:

$$\bar{Y}^*_{spt} = \sum_{k=-3}^{2} \beta_k \tau_k + \varphi_t + \varepsilon_{spt} \tag{7}$$

τ_k are event time indicators, which are set to 1 if year t is k years from a principal transition and φ_t are year fixed effects. Standard errors are clustered by school-spell to allow for the

[18] Due to panel length limitations and the relatively few principals who remain in the same school in Oregon for 6 or more years, we conduct the same analysis in Oregon within a range of four years (2 years pre and 2 years post principal transition).

correlation of errors over time within each unique event.

These event studies may not clearly indicate the impact of principal quality (induced by the compositional effect of the principal transition) because school performance in the pre-period reflects time-varying school factors that principals cannot control. Thus, we also estimate event studies for a set of comparison schools. These are schools that are in the top or bottom quartile using the same three-year moving average as above but did not change principals in the following year. More specifically, we restrict our comparison group to schools that had the same principal across the entire event window.[19] This provides a reasonable counterfactual for schools that changed principals.[20]

All else equal, we anticipate that low-performing schools that change principals will see an increase in student test scores under the new principal (and vice-versa for high-performing schools). Given the similarity of r_x^{SamePrin} and r_x^{DiffPrin} in Figure 2, however, we expect that comparison schools that keep their principal will also experience these same changes.

Our principal-switching analyses show that schools' performance follows a similar mean reversion pattern, irrespective of whether their principals stay or leave. We plot the results of Equation 7 in Figure 4. With the exception of high-performing NYC schools in math, the overall patterns are consistent in demonstrating that the observed changes in the post-period are not driven by compositional effects of changing principals. Instead, schools are following a dynamic, mean-reverting pattern that creates the *illusion* of principal effects. We formalize these results in a series of classic 2×2 pre/post difference-in-difference estimates, including an indicator for schools that changed principals (see Appendix Tables A.9–A.11). Outside of

[19] Note that unless a principal stayed exactly six years, there are multiple six-year event windows we could choose for comparison schools. To avoid duplication, we impose a restriction whereby a school-by-principal spell can only be used once as a comparison school. We then choose the six-year window that includes the most recent years of data possible.

[20] One challenge with these comparisons, however, is that the principal transition itself may have a disruptive effect on school performance in addition to the compositional effect of the change in leadership quality (Bartanen, Grissom, and Rogers 2019). Prior work shows that there is an "Ashenfelter dip" in school performance in the year prior to a principal transition (Bartanen, Grissom, and Rogers 2019; Miller 2013; Laing et al. 2016) and that turnover has a small negative impact on performance under the new principal. Accordingly, we view these event studies with some caution, though these disruptive turnover dynamics tend to be small in magnitude and will not affect the comparison schools that kept the same principal.

high-performing NYC schools in math, we fail to reject the null hypothesis in 38 of 38 tests (at a 95% confidence level) that there is no difference in performance trajectories between schools that did and did not change principals.[21]

5.5 Variance Decomposition with Autocorrelated Errors

Based on this deeper understanding of the within-school correlation of residuals, we return to our variance decomposition models, which attempted to establish the magnitude of principals' contributions to student test scores.

In parsing the variance of school performance between schools and principals, the baseline models in Table 1 assume that the residuals (school-by-year cells nested within principals) are independent and identically distributed—an assumption that is clearly violated based on the patterns in Figure 2. This violation leads to a bias in the magnitude of the estimated variance components. In particular, the principal variance component is biased upwards. Because the typical principal remains in a school for a short period (2–5 years), year-to-year fluctuations that are unrelated to principal effectiveness constitute a non-trivial portion of the principal's estimated effect. By directly estimating the error variance, the mixed model adjusts for yearly fluctuations and avoids inflating the principal variance component. In the case of positive autocorrelation, however, the adjustment is insufficient and the principal variance component remains inflated.

To address this bias, we re-estimate our variance components models with an autoregressive error structure, where the correlation is an additional parameter to be estimated along with the random effect variances. In particular, if we assume the error term in Equation 6 (ϵ_{spt}) follows a first-order autoregressive structure:

[21] For completeness, we provide a full set of event study plots for all four models in math and reading in Appendix Figures A.9–A.15.

$$
\begin{aligned}
\epsilon_{spt} &= \rho \epsilon_{sp,t-1} + \nu_{spt} \\
\text{where } \nu_{spt} &\sim N(0, \sigma^2)
\end{aligned}
\tag{8}
$$

then, we can re-write Equation 6 with the composite error term:

$$
\bar{Y}^*_{spt} = \theta_s + \theta_{s,p} + [\rho \epsilon_{sp,t-1} + \nu_{spt}] \tag{9}
$$

where we directly estimate the AR(1) term ρ along with the school, principal, and residual variance components. Correctly modeling the positive autocorrelation structure will increase the estimated variance component of the residual and shrink the principal variance component, producing a more accurate estimate of the magnitude of principals' effects.

Once we appropriately model the semi-persistent ebbs and flows in school performance—variation that should not be attributed to principal effects—our estimates of the magnitude of principal effects on student test scores and attendance are, in essence, zero. We show results from AR(2) models in Table 2, which we found fit better than AR(1) models.[22] We find that while the estimated school-level variance components are essentially unchanged relative to Table 1, the principal-level variance components are effectively zero.

The issues we raise here can also be described in terms of the identification assumptions required for interpreting principal VA as causal effect estimates. In the typical school fixed effects approach, internal validity hinges on an assumption that the sources of unobserved school heterogeneity—e.g., the school's neighborhood, the characteristics of successive cohorts of students—are time invariant. Our analyses clearly show that this assumption is violated. This introduces bias into each of the individual principal VA estimates and leads to an overstatement of the magnitude of principal effects. Once we properly account for the correlation structure of time-varying, unobserved school-level heterogeneity, we find little evidence that variation in principal effectiveness contributes systematically to changes in

[22] AR(1) models as well as results for reading and attendance are shown in Appendix Tables A.2–A.3 and A.12–A.13.

student test scores.

5.6 What Drives Changes in School Performance?

To this point, our analysis establishes that there is substantial within-school variation in student test score performance. These changes over time exhibit autocorrelation, but this dynamic pattern does not seem to be driven by principals. What, then, might be driving the dynamic component? We examine two sets of mechanisms. The first are fluctuations in teacher composition. While principals are perceived as key human capital managers for schools—both through considerable autonomy to hire new teachers and their influence over teacher retention—they undoubtedly have incomplete control over teacher composition. For instance, a new principal typically inherits most of the teachers hired under their predecessor(s). Additionally, while principals may have autonomy in choosing which teacher applicants to hire, the applicant pool may be mostly out of their control. Finally, while prior work suggests that effective principals can lower teacher turnover, teachers' mobility decisions can not be fully attributed to principals. In short, a school's teaching staff is constantly in flux and likely only partially attributable to the principal.

Changes in teacher composition likely contribute to the autocorrelated nature of school performance. To see this, consider the retirement of a highly effective teacher. The subsequent impact on school performance will manifest as both a short-term disruptive effect *and* a longer-term compositional effect. The latter effect, which is the difference between the effectiveness of the retiring teacher and their replacement, is not a one-time shock to school performance because the replacement likely stays in the school for multiple years.

To investigate this, we examine r_x^{SamePrin} and r_x^{DiffPrin} for two measures of teacher composition: mean teacher experience and mean teacher value-added. For the latter, we follow the methodology of (Chetty, Friedman, and Rockoff 2014a) to construct a drift-adjusted VA measure, but we use only a teacher's performance from a different school.[23] This avoids

[23] This is functionally similar to how (Chetty, Friedman, and Rockoff 2014a) construct leave-out VA for

estimating teacher VA from the same test score residuals that form our school performance measures. Figure 5 shows the results. The logic is parallel to the results for test score residuals in Figure 2. If principals drive systematic changes in teacher composition, we should observe stronger correlations within the same principal than across principals.

This is not the case. For both measures, there is a clear decreasing pattern in the correlations over time, demonstrating ebbs and flows in a school's average teacher composition. These changes, however, do not appear to be driven by principals. These results provide some insight regarding the school performance patterns. Abundant evidence demonstrates that teacher quality is a key within-school factor for student test-score growth and longer-term outcomes (e.g., Chetty, Friedman, and Rockoff 2014b). Changes in a school's teaching staff, then, will drive changes in $\bar{Y}^{*}_{spt}$.[24] To the extent that principals can meaningfully shape teacher quality through strategic human capital management, this would be a key mechanism through which they drive changes in student test scores. These results challenge that chain of logic. This is, of course, not to say that principals are unable to shape students outcomes through human capital management in certain contexts or in schools with alternate governance structures. We do not, however, see evidence that this is occurring systematically across these contexts, at least with respect to teacher VA and experience. While this may be explained by constraints (e.g., incomplete autonomy to dismiss teachers or a persistently weak hiring pool), it could also reflect that principals' preferences with respect to teacher composition are not aligned with these measures (e.g., Ballou 1996; Goldring et al. 2015).[25]

The second mechanism we investigate is the composition of entering cohorts of students

their quasi-experimental test of teacher switches. Whereas they predict teacher VA in year t using test score residuals from more than three years prior or two years after, we predict teacher VA in year t using all test score residuals where the teacher was working in a different school.

[24] Appendix Tables A.14–A.16 provide empirical support for this claim. Specifically, we estimate via first differences the relationship between school performance and mean teacher VA (using the leave-out-current-school measure and experience). We find that changes in VA positively predict changes in test score residuals, except for reading in NYC. In all contexts, the relationship between change in mean teacher experience and change in test score residuals is close to zero. This is unsurprising, perhaps, because we are looking at all of the teachers in the school, not just those teaching students whose test scores contribute to the school performance measure.

[25] Even if principals seek to maximize teacher VA, there may be an informational constraint at the time of hire about who will be a high-VA teacher.

over time. The intuition here is that different student cohorts entering from common "feeder" schools and neighborhoods may have different performance profiles as the effectiveness of students' prior schools ebbs and flows or compositional changes occur in the neighborhood. These student cohorts remain in the school for several years and so their entry "shock" is not purely transient. We anticipate that principals have relatively little influence over students' prior-school outcomes. However, if students' prior-school outcomes follow the same autocorrelation pattern as their current school performance (higher correlations in the most proximal years, declining over time), it would suggest that the dynamic patterns in school performance may be influenced by these trends.

We document in Figure 6 that students' prior-school lagged test score outcomes follow the same dynamic pattern as their performance in their current school, and therefore may explain some of the semi-persistent variation in school performance.[26] In particular, students' prior-school outcomes are strongly correlated in immediately adjacent years, but these autocorrelations tail off over the subsequent years. These results highlight the dynamic, shifting nature of the prior performance of students entering into schools and offers at least suggestive evidence of what may be driving autocorrelations in school performance. While VA models that control for prior test scores will help account for drift in incoming cohort quality, the presence of these patterns hints at the possibility of cohort-level fluctuations on unobservables, which would contribute to the dynamic nature of school performance.

5.7 Analyzing School-Switching Principals

As an additional check on our main results, we leverage principals whom we observe leading multiple schools. Here, our aim is to compare the stability of school performance within principals for years when they worked in the same school versus a different school. In particular, we are interested in the different school correlations. Drawing on the framework

[26] Note that we cannot produce these results for NYC, since we do not have information about when a student first enrolls in a school.

in Equation 2, this correlation is:

$$r_x^{\text{DiffSch}} = \frac{\sigma^2_{\delta^F} + \sigma_{\delta^D_x} + \sigma_{\mu^F_x}}{\sigma^2_Y}$$
$$\text{where } \sigma_{\mu^F_x} = \text{cov}(\mu^F_{j(p,t)}, \mu^F_{k(p,t+x)}) < \sigma^2_{\mu^F} \tag{10}$$

Given our interpretation of the main results that principals are not driving persistent changes in school performance, we expect that r_x^{DiffSch} will be small in magnitude. As with r_x^{DiffPrin} in Equation 4, however, part of r_x^{DiffSch} reflects the possibility of principal sorting. If principals tend to transfer to schools that are similar (different) in terms of fixed factors μ^F that affect student test score performance, r_x^{DiffSch} will be higher (lower).[27]

Figure 7 plots r_x^{SameSch} (which is the same as r_x^{SamePrin} from the main results) and r_x^{DiffSch}. Consistent with our expectations, r_x^{DiffSch} is small in magnitude and substantially smaller than r_x^{SameSch}, hovering around only 0.2. Still, could this small correlation across different schools—which is relatively stable over time—suggest a small contribution of principal effectiveness to school performance? This is unlikely—if $\sigma^2_{\delta^F} > 0$, we should have seen greater separation between r_x^{SamePrin} and r_x^{DiffPrin} in Figure 2. Instead, this correlation likely reflects sorting. To show this, we can estimate r_x^{DiffSch} for our teacher and student composition measures. As demonstrated by Appendix Figures A.16 and A.17, we observe positive correlations of roughly the same magnitude, reinforcing that the small amount of stability in school performance observed within principals across different schools is not indicative of principals' effects on student outcomes, but rather of principals sorting to similar school environments.

[27] As with teachers, the principal labor market tends to be highly localized. Prior work in Tennessee, for instance, finds that nearly all transferring principals remain in the same district (Grissom and Bartanen 2019a).

6 Conclusion

Principals play a central role in schools, and there is substantial interest from researchers and policy makers in understanding the extent to which principals affect student outcomes, including test scores. This interest stems, at least in part, from a dominant paradigm asserting that effective principals should produce better outcomes for students. Even without calculating explicit measures of principal value-added, nearly all states now incorporate student outcomes-based measures into principal evaluation systems. There has also been a proliferation of research using test score-based measures to draw inferences about the effectiveness of policies and practices related to school leadership.

Our key empirical result is that most or all of the within-school variation in school performance—as measured by the average student's yearly test score residual—does not appear to be caused by effectiveness differences across principals. Specifically, while we find meaningful within-school variation in student test score performance when comparing across principals (which is the identifying variation for principal VA models), this variation is driven by transient school factors that are likely to have occurred regardless of who was leading the school. Because these school factors exhibit some persistence across years, they create the illusion of principal effects even when applying shrinkage approaches that assume no serial correlation of residuals.

From this empirical result come two important findings. The first is that existing value-added approaches that use student outcomes to measure principal performance or quality are flawed. By misattributing to principals the effects of dynamic changes in school-level factors that principals do not appear to control, these value-added approaches yield biased estimates of principal effects. Given the short tenure of the typical principal, most of their "value added" reflects the (mis)fortune of when they entered the school, as opposed to their own leadership effectiveness. To the extent that the current test score performance of a principal's school informs high-stakes decision-making (contract renewals, salary increases, etc.), our results imply substantial inefficiencies.

The second finding concerns the magnitude of principals' effects on student outcomes. Once accounting for the dynamic nature of school performance, variation in principal quality explains little to none of the observed variation in student test scores. This finding directly refutes the existing literature, where principals' estimated impacts—based on value-added methodology—are relatively large in magnitude. Given that many principals do not remain in the same school beyond a few years, we suggest some caution in the interpretation of this finding. The frequent churn of school leaders could be part of the substantive explanation for why we observe little variation in principals' measured effects. Nonetheless, our key results largely hold even among longer-tenured principals and in a context (New York City) where relatively more principals remain in their schools for an extended period.

It is also important to note that we do not establish that principals *cannot* meaningfully influence the school factors that drive student achievement. For instance, there are examples of contexts where principals have the information and/or autonomy required to engage in strategic human capital management (e.g., Jacob 2011; Grissom and Bartanen 2019b; Boyd et al. 2011; Goldring et al. 2015). Our results, however, suggest that such behaviors are not driving systematic differences in school performance across principals, on a large scale.

In considering what might explain these results, it seems unlikely that principals seek to maximize objectives that are entirely orthogonal to improving student achievement and attendance, particularly given that our panels overlap with the height of the accountability movement in U.S. education policy. Thus, one potential explanation is that the typical principal faces considerable constraints on their ability to shape school factors like teacher composition or skill, particularly in the short run. Another possibility is that principals focus on additional goals beyond raising average test scores or attendance, such as student and teacher well-being. The latter could follow from the former if principals internalize their limited capacity to drive test score gains. In this vein, principals may contribute more substantially to other important conditions for teaching and learning in a school or to longer-term outcomes that contemporaneous test scores and attendance fail to measure.

A key policy implication is that states and districts should not rely heavily on student outcomes (namely, student test scores or attendance) in forming their judgments of leadership effectiveness. Although this may seem antithetical to the view that improving student outcomes is a central goal for schools, we stress that holding principals accountable for these outcomes likely leads to substantial misclassification errors that may further undermine school performance. While not an explicit focus of this analysis, we found that ratings of principals' performance from their supervisors and survey-based measures from teachers were able to differentiate principals. Several districts currently supplement this evidence on school quality with surveys of students and families. There is no guarantee regarding the validity of these measures, but given our findings that test-score-based measures do not reflect real differences in principal quality, they are almost certainly a more useful tool.

Finally, we urge additional study—particularly using designs that credibly support causal inferences—of the effects of differences in principal behaviors and skills on student near- and longer-term outcomes that would allow researchers and practitioners to look inside the "black box" of effective leadership. Given the inherent challenges in measuring principal performance, a deeper understanding of the mechanisms that link effective leadership to student outcomes is an important avenue for future research.

References

Bacher-Hicks, Andrew, Stephen B Billings, and David J Deming. 2019. *The School to Prison Pipeline: Long-run Impacts of School Suspensions on Adult Crime.* Technical report, NBER Working Paper Series 26257. National Bureau of Economic Research.

Ballou, Dale. 1996. “Do Public Schools Hire the Best Applicants?” *The Quarterly Journal of Economics* 111, no. 1 (February): 97–133.

Bartanen, Brendan. 2020. “Principal Quality and Student Attendance.” *Educational Researcher* 49 (2): 101–113.

Bartanen, Brendan, Jason A Grissom, and Laura K Rogers. 2019. “The Impacts of Principal Turnover.” *Educational Evaluation and Policy Analysis* 41 (3): 350–374.

Bartanen, Brendan, and Aliza N. Husain. 2021. *Connected Networks in Principal Value-Added Models.* Technical report, EdWorkingPaper 21-397. Providence, RI: Annenberg Institute at Brown University.

Boyd, D., P. Grossman, M. Ing, H. Lankford, S. Loeb, and J. Wyckoff. 2011. “The Influence of School Administrators on Teacher Retention Decisions.” *American Educational Research Journal* 48 (2): 303–333.

Branch, Gregory, Eric A. Hanushek, and Steven Rivkin. 2012. *Estimating the Effect of Leaders on Public Sector Productivity: The Case of School Principals.* Technical report, NBER Working Paper No. 17803. Cambridge, MA: National Bureau of Economic Research.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014a. “Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates.” *American Economic Review* 104 (9): 2593–2632.

———. 2014b. “Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood.” *American Economic Review* 104 (9): 2633–2679. ISSN: 0002-8282.

Chiang, Hanley, Stephen Lipscomb, and Brian Gill. 2016. “Is School Value Added Indicative of Principal Quality?” *Education Finance and Policy* 11 (3): 283–309.

Coelli, Michael, and David A. Green. 2012. “Leadership Effects: School Principals and Student Outcomes.” *Economics of Education Review* 31 (1): 92–109.

Cullen, Julie Berry, Eric A. Hanushek, Gregory Phelan, and Steven G. Rivkin. 2021. “Performance Information and Personnel Decisions in the Public Sector: The Case of School Principals.” *Journal of Human Resources,* no. May, 0619–10272R1.

Dhuey, Elizabeth, and Justin Smith. 2014. “How Important Are School Principals in the Production of Student Achievement?” *Canadian Journal of Economics* 47 (2): 634–663.

———. 2018. “How School Principals Influence Student Learning.” *Empirical Economics* 54 (2): 851–882.

Donaldson, Morgaen, Madeline Mavrogordato, Shaun M. Dougherty, Reem Al Ghanem, and Peter Youngs. 2021. “Principal Evaluation Under the Elementary and Secondary Every Student Succeeds Act: A Comprehensive Policy Review.” *Education Finance and Policy* 16 (2): 347–361.

Gates, Susan, Matthew Baird, Benjamin Master, and Emilio Chavez-Herrerias. 2019. *Principal Pipelines: A Feasible, Affordable, and Effective Way for Districts to Improve Schools.* Santa Monica, CA: RAND Corporation.

Goldhaber, Dan, and Michael Hansen. 2013. “Is it Just a Bad Class? Assessing the Long-term Stability of Estimated Teacher Performance.” *Economica* 80 (319): 589–612.

Goldring, Ellen, Jason A. Grissom, Mollie Rubin, Christine M Neumerski, Marisa Cannata, Timothy Drake, and Patrick Schuermann. 2015. “Make Room Value Added: Principals’ Human Capital Decisions and the Emergence of Teacher Observation Data.” *Educational Researcher* 44 (2): 96–104.

Grissom, Jason A., and Brendan Bartanen. 2019a. “Principal Effectiveness and Principal Turnover.” *Education Finance and Policy* 14, no. 3 (July): 355–382.

———. 2019b. “Strategic Retention: Principal Effectiveness and Teacher Turnover in Multiple-Measure Teacher Evaluation Systems.” *American Educational Research Journal* 56 (2): 514–555.

Grissom, Jason A., Anna J Egalite, and Constance A Lindsay. 2021. *How Principals Affect Students and Schools: A Systematic Synthesis of Two Decades of Research.* Technical report. New York: The Wallace Foundation.

Grissom, Jason A., Demetra Kalogrides, and Susanna Loeb. 2015. “Using Student Test Scores to Measure Principal Performance.” *Educational Evaluation and Policy Analysis* 37 (1): 3–28.

Grissom, Jason A., and Susanna Loeb. 2011. “Triangulating Principal Effectiveness.” *American Educational Research Journal* 48, no. 5 (October): 1091–1123.

Hanushek, Eric. 1971. “Teacher Characteristics and Gains in Student Achievement: Estimation Using Micro Data.” *American Economic Review: Papers and Proceedings* 61 (2): 280–288.

Jacob, Brian A. 2011. “Do Principals Fire the Worst Teachers?” *Educational Evaluation and Policy Analysis* 33 (January): 403–434.

Jacob, Brian A., and L. Lefgren. 2007. “What Do Parents Value in Education? An Empirical Investigation of Parents’ Revealed Preferences for Teachers.” *The Quarterly Journal of Economics* 122, no. 4 (November): 1603–1637.

Koedel, Cory, Kata Mihaly, and Jonah E. Rockoff. 2015. “Value-Added Modeling: A Review.” *Economics of Education Review* 47 (August): 180–195.

Laing, Derek, Steven Rivkin, Jeffrey Schiman, and Jason Ward. 2016. *Decentralized Governance and the Quality of School Leadership.* Technical report, NBER Working Paper 22061. Cambridge, MA: National Bureau of Economic Research.

Liebowitz, David D., and Lorna Porter. 2019. "The Effect of Principal Behaviors on Student, Teacher, and School Outcomes: A Systematic Review and Meta-analysis of the Empirical Literature." *Review of Educational Research* 89 (5): 785–827.

Miller, Ashley. 2013. "Principal turnover and student achievement." *Economics of Education Review* 36:60–72.

Murnane, Richard J. 1975. *The Impact of School Resources on the Learning of Inner City Children.* Cambridge, MA: Balinger Publishing Company.

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

Rubin, D. B., E. A. Stuart, and E. L. Zanutto. 2004. "A Potential Outcomes View of Value-Added Assessment in Education." *Journal of Educational and Behavioral Statistics* 29 (1): 103–116.

Sorensen, Lucy C., Shawn D. Bushway, and Elizabeth J. Gifford. 2021. "Getting Tough? The Effects of Discretionary Principal Discipline on Student Outcomes." *Education Finance and Policy,* 1–74.

Steinberg, Matthew P., and Lauren Sartain. 2015. "Does Teacher Evaluation Improve School Performance? Experimental Evidence from Chicago's Excellence in Teaching Project." *Education Finance and Policy* 10 (4): 535–572.

Steinberg, Matthew P., and Haisheng Yang. 2022. "Does Principal Professional Development Improve Schooling Outcomes? Evidence from Pennsylvania's Inspired Leadership Induction Program." *Journal of Research on Educational Effectiveness.*

Taylor, Eric S., and John H. Tyler. 2012. "The Effect of Evaluation on Teacher Performance." *The American Economic Review* 102 (7): 3628–3651.

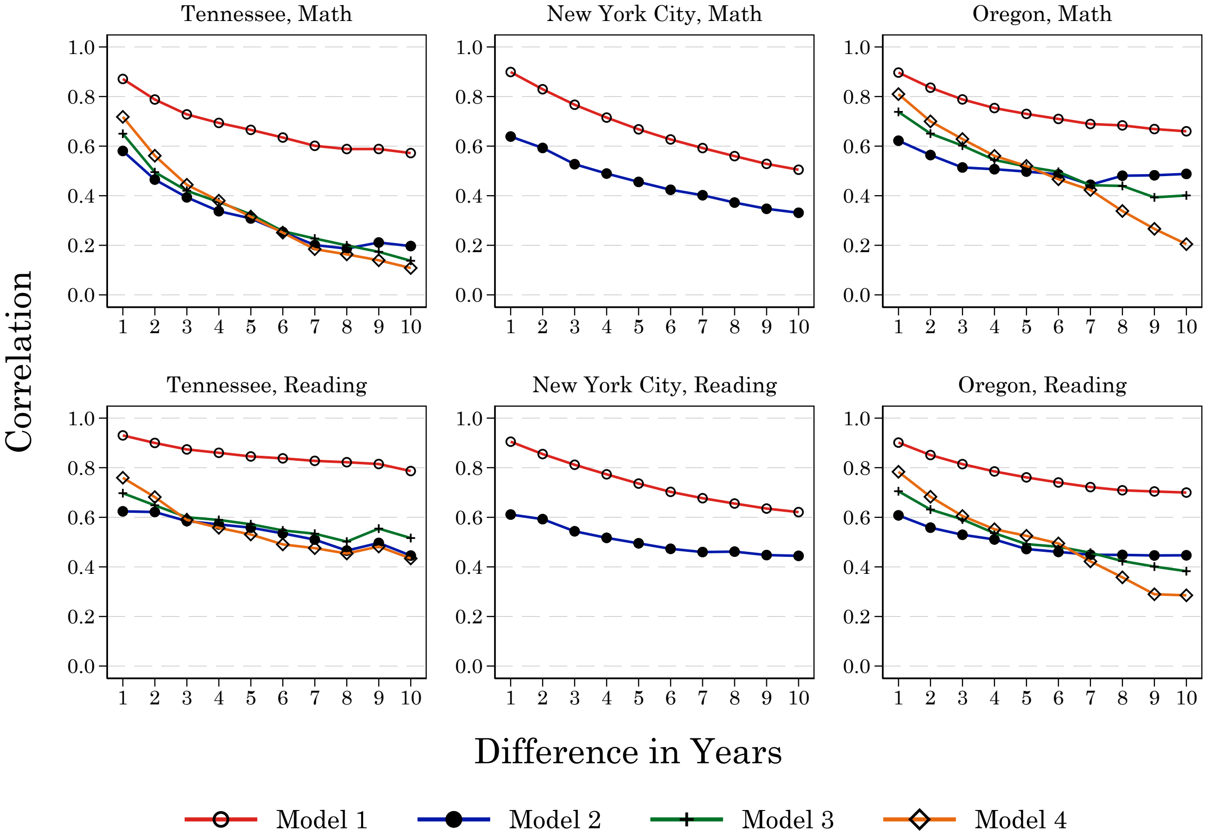


Figure 1: Autocorrelation Vectors

Notes: Figures report autocorrelation "drift" vectors generated from sample-size-precision-weighted correlations of school-by-year mean residualized test score generated from Equation 5. Correlations are between year t and $t + x$ for the same school, where x is denoted by the x-axis value. Model 1 adjusts for student demographic characteristics, Model 2 adds cubic polynomials of lagged-test scores and attendance, Model 3 restricts to students in first year in school, Model 4 uses prior-school outcomes. Models 3 and 4 not estimated for NYC because we do not observe year of first enrollment. Sample sizes for each correlation are shown in Appendix Table A.4.

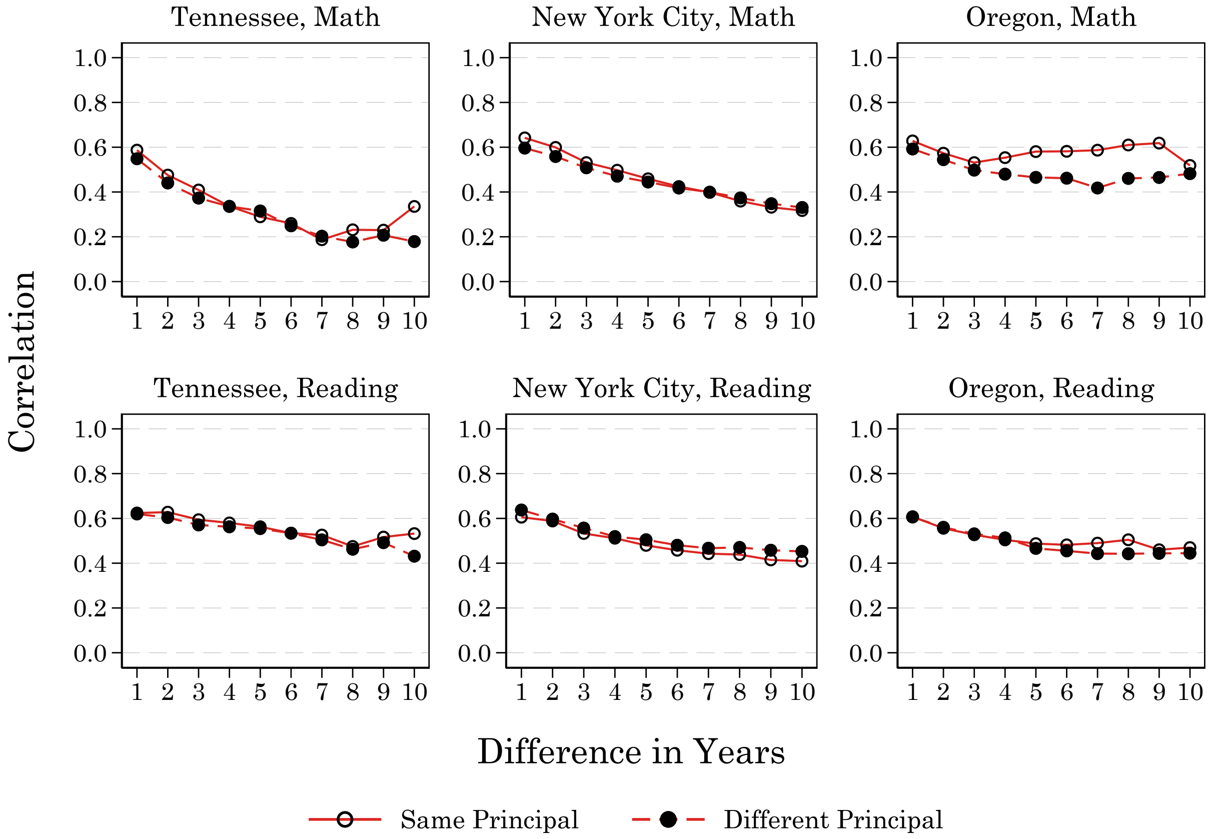
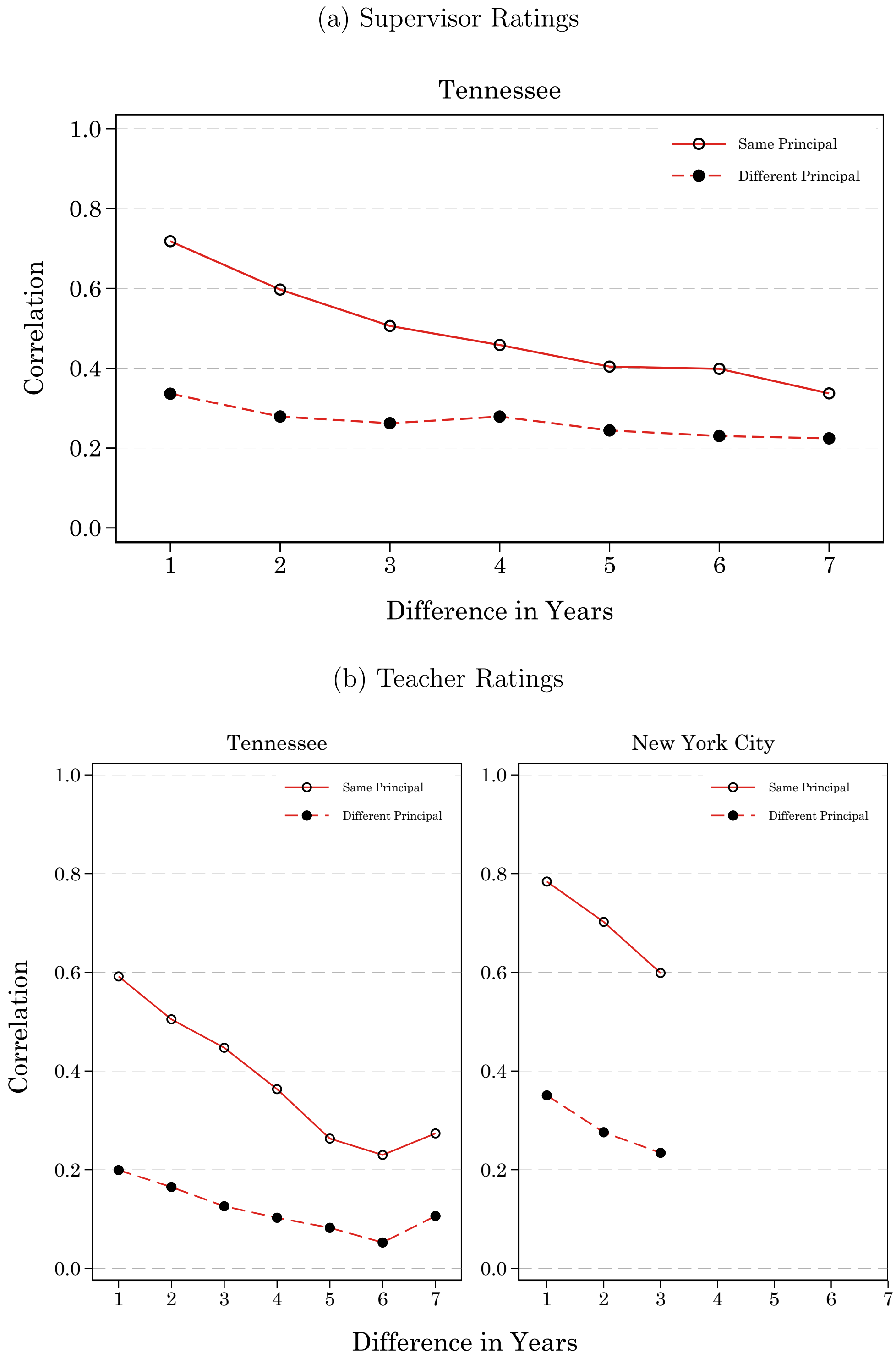


Figure 2: Autocorrelations Within and Between Principals

Notes: Figures report autocorrelation "drift" vectors generated from sample-size-precision-weighted correlations of school-by-year mean residualized test score generated from Equation 5. Correlations are between year t and $t + x$ for the same school, where x is denoted by the x-axis value. Same principal denotes the sub-sample of school-by-year pairs where the principal is the same in both years. Different principal denotes the sub-sample where the principal in year t is different than year $t + x$. Residualization models adjust for student demographic characteristics and cubic polynomials of lagged-test scores and attendance (Model 2). Sample sizes for each correlation are shown in Appendix Table A.4.


(a) Supervisor Ratings

(b) Teacher Ratings

Figure 3: Autocorrelations of Perceptions of Performance, Within and Between Principals

Notes: Figures report autocorrelation "drift" vectors generated from supervisor and teacher ratings. Correlations are between year t and $t+x$ for the same school, where x is denoted by the x-axis value. Same principal denotes the sub-sample of school-by-year pairs where the principal is the same in both years. Different principal denotes the sub-sample where the principal in year t is different than year $t + x$. Correlations are unweighted for supervisor ratings. For teacher ratings, we weight by the number of teachers that responded to the survey from which the measure is constructed.

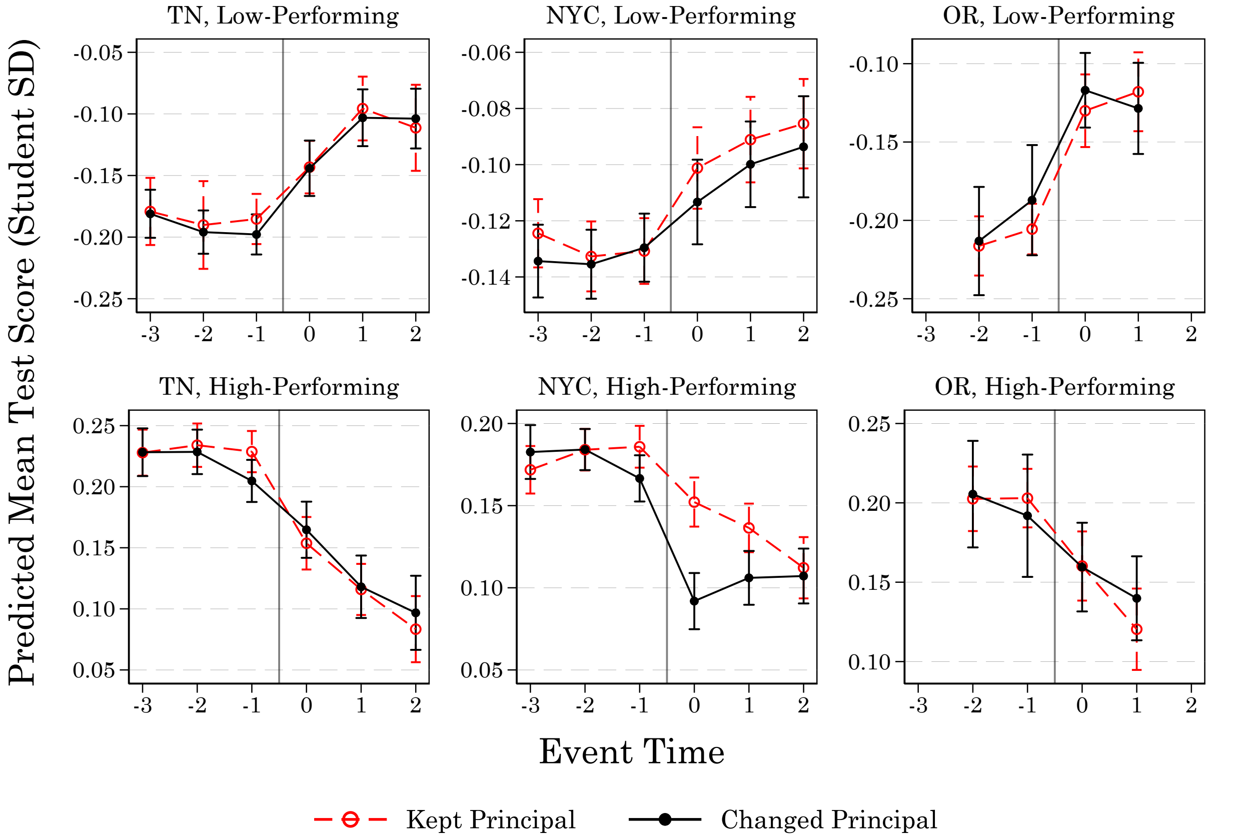


Figure 4: Event Study (Math)

Notes: Figures report event-study estimates and 95% confidence intervals from Equation 7. Residualization models adjust for student demographic characteristics and cubic polynomials of lagged-test scores and attendance (Model 2). High-performing and low-performing defined by top- and bottom-quartile three-year moving average of school-by-year mean test score residuals. Due to panel length limitations in Oregon, we restrict the event time to a range of four years. Standard errors adjusted for clustering at the school-spell level. School-by-year cells are weighted by the number of students contributing to the mean test score residual measure in the given year.

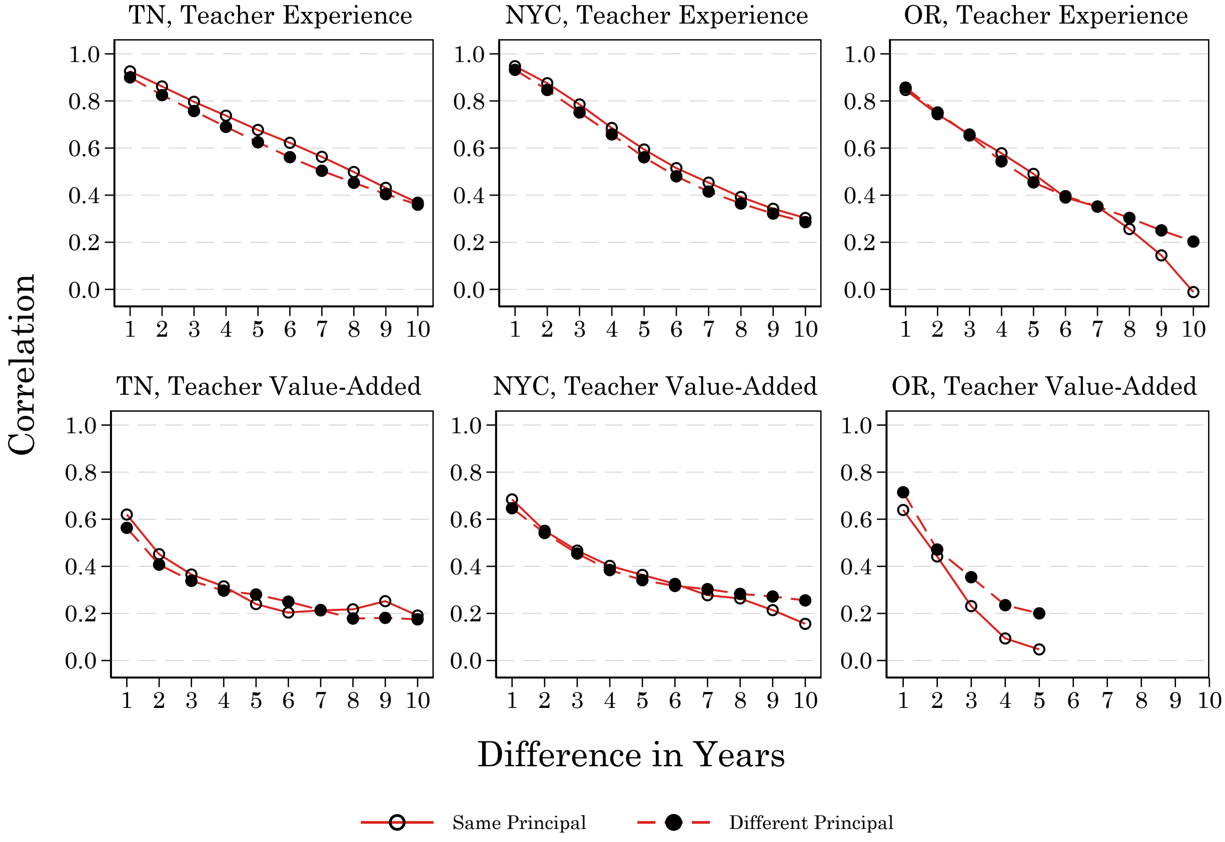


Figure 5: Autocorrelation Vectors for Teacher Composition

Notes: Figures report autocorrelation "drift" vectors generated from sample-size-precision-weighted correlations of school-by-year mean teacher experience and value-added (pooling math and reading teachers). Correlations are between year t and $t + x$ for the same school, where x is denoted by the x-axis value. Same principal denotes the sub-sample of school-by-year pairs where the principal is the same in both years. Different principal denotes the sub-sample where the principal in year t is different than year $t + x$. For teacher experience, school-by-year cells are weighted by the number of teachers in the school. For VA, school-by-year cells are weighted by the number of teachers with a VA estimate.

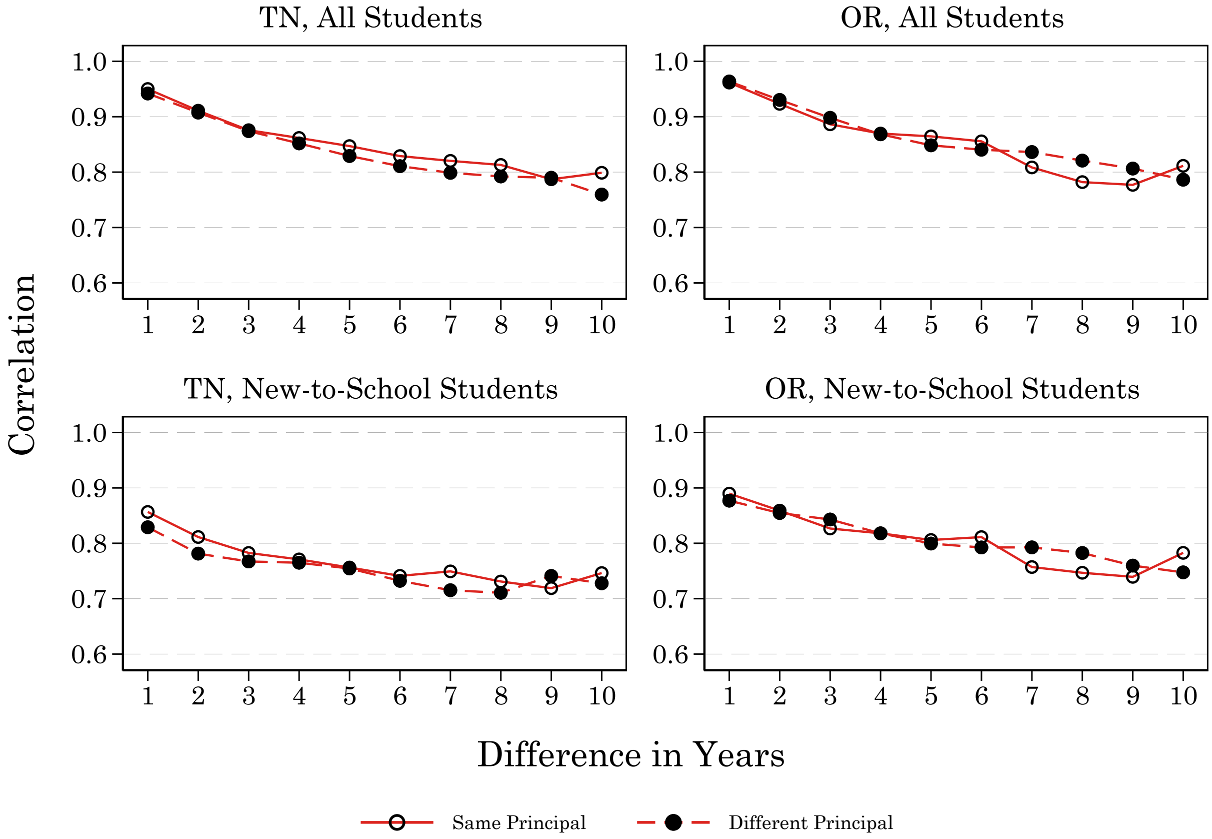


Figure 6: Autocorrelation Vectors for Student Composition Using Prior-School Lagged Test Scores

Notes: Figures report autocorrelation "drift" vectors generated from sample-size-precision-weighted correlations of school-by-year prior-school outcomes for both new-to-school and all students generated from Equation 5. Correlations are between year t and $t + x$ for the same school, where x is denoted by the x-axis value. Same principal denotes the sub-sample of school-by-year pairs where the principal is the same in both years. Different principal denotes the sub-sample where the principal in year t is different than year $t + x$. OR results exclude HS students as very few new-to-school students have a prior-school *and* a current-year score because only a small number of 9th-grade students appear in our sample prior to 2014 and none afterwards.

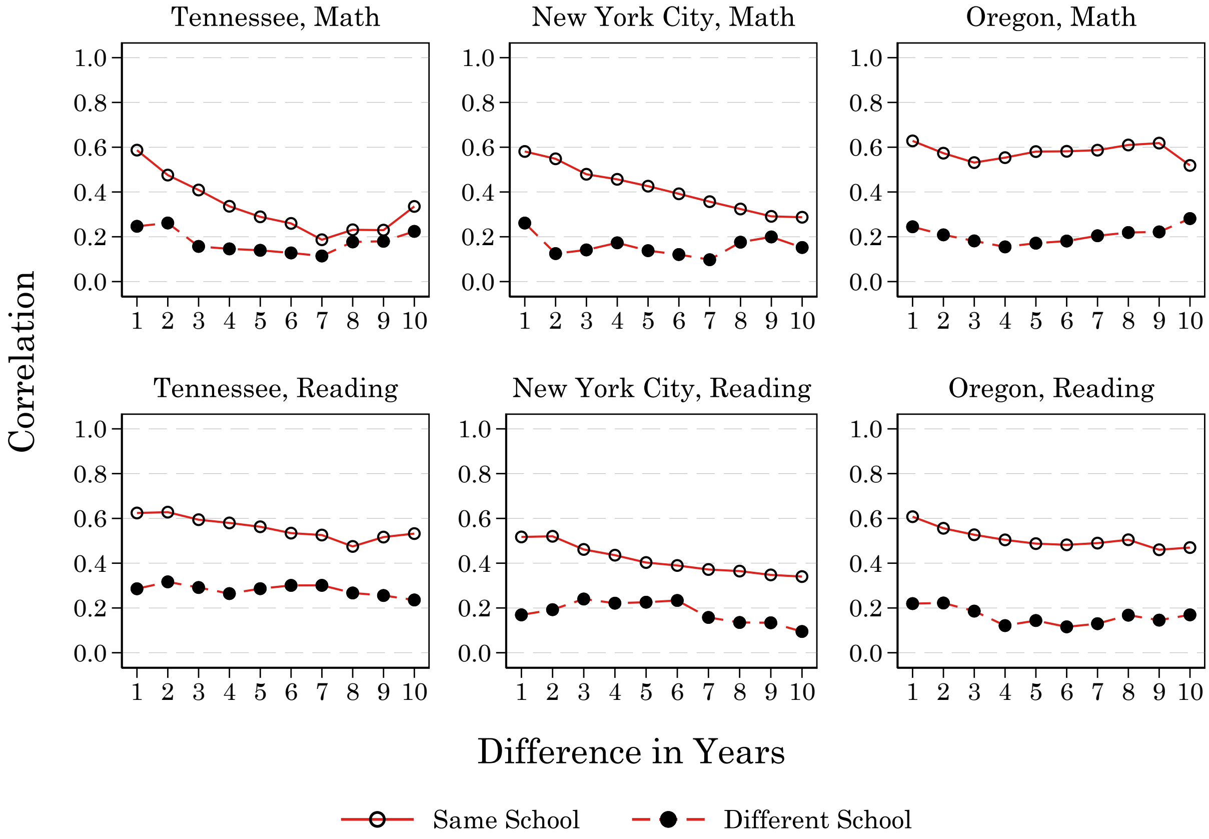


Figure 7: Autocorrelations Within and Between Schools

Notes: Figures report autocorrelation "drift" vectors generated from sample-size-precision-weighted correlations of principal-by-year mean residualized test score generated from Equation 5. Correlations are between year t and $t + x$ for the same principal, where x is denoted by the x-axis value. Same same denotes the sub-sample of principal-by-year pairs where the school is the same in both years. Different school denotes the sub-sample where the school in year t is different than year $t + x$. Residualization models adjust for student demographic characteristics and cubic polynomials of lagged-test scores and attendance (Model 2).

Table 1: Variance Decomposition (Math)

	Model 0	Model 1	Model 2	Model 3	Model 4
Panel A: Tennessee					
Random Effects Parameters (SD)					
School	0.411	0.296	0.129	0.169	0.166
Principal	0.136	0.133	0.090	0.108	0.116
Residual	0.173	0.168	0.154	0.162	0.147
Variance Components (%)					
School	77.7	65.6	34.2	43.0	44.0
Principal	8.5	13.2	16.6	17.4	21.6
Residual	13.8	21.2	49.1	39.6	34.3
N (Schools)	1841	1841	1841	1421	1300
N (Principal-by-School)	4797	4797	4797	3265	3007
N (School-by-Year Cells)	17553	17553	17553	10461	9876
Mean Students Per Cell	291	291	291	163	313
Panel B: New York City					
Random Effects Parameters (SD)					
School	0.500	0.225	0.111		
Principal	0.170	0.124	0.060		
Residual	0.150	0.139	0.113		
Variance Components (%)					
School	83.0	59.3	43.2		
Principal	9.6	18.1	12.5		
Residual	7.5	22.6	44.3		
N (Schools)	1317	1317	1317		
N (Principal-by-School)	3489	3489	3489		
N (School-by-Year Cells)	18350	18350	18350		
Mean Students Per Cell	338	338	338		
Panel C: Oregon					
Random Effects Parameters (SD)					
School	0.362	0.310	0.149	0.212	0.198
Principal	0.117	0.111	0.071	0.095	0.093
Residual	0.143	0.138	0.138	0.146	0.128
Variance Components (%)					
School	79.5	75.5	48.0	59.7	60.9
Principal	8.2	9.7	11.0	11.9	13.5
Residual	12.3	14.8	41.0	28.4	25.6
N (Schools)	1269	1269	1269	863	816
N (Principal-by-School)	3559	3559	3559	1869	1655
N (School-by-Year Cells)	11815	11815	11815	5034	4322
Mean Students Per Cell	243	243	243	145	229
Student Characteristics		✓	✓	✓	✓
Prior-Year Test Scores			✓	✓	
New-to-School Students Only				✓	
Prior-School Test Scores					✓

Notes: Cells report standard deviations of variance components and percentage of overall variance explained from Equation 6. Model 0 uses raw test scores, Model 1 adjusts for student demographic characteristics, Model 2 adds cubic polynomials of lagged-test scores and attendance, Model 3 restricts to students in first year in school, Model 4 uses prior-school outcomes. Models 3 and 4 not estimated for NYC because we do not observe year of first enrollment. All models include grade and year fixed effects. Demographic covariates include prior grade retention, gender, race/ethnicity, disability status, 504 plan designation, participation in migrant or Indian education program and the school averages of the preceding characteristics. All samples restricted to observations with at least 25 students in each school-by-year cell.

Table 2: Variance Decomposition Results with Autocorrelated Errors (Math)

	Model 0	Model 1	Model 2	Model 3	Model 4
Panel A: Tennessee					
Random Effects Parameters (SD)					
School	0.411	0.297	0.129	0.171	0.166
Principal	0.000	0.000	0.000	0.000	0.000
Residual	0.216	0.211	0.177	0.193	0.185
AR(2) Parameters					
Correlation $(t-1)$	0.567	0.571	0.350	0.423	0.508
Correlation $(t-2)$	0.045	0.045	0.113	0.084	0.093
Panel B: New York City					
Random Effects Parameters (SD)					
School	0.499	0.223	0.111		
Principal	0.000	0.000	0.009		
Residual	0.222	0.184	0.126		
AR(2) Parameters					
Correlation $(t-1)$	0.744	0.661	0.272		
Correlation $(t-2)$	0.062	0.076	0.193		
Panel C: Oregon					
Random Effects Parameters (SD)					
School	0.362	0.310	0.150	0.212	0.197
Principal	0.000	0.000	0.030	0.000	0.000
Residual	0.183	0.175	0.152	0.174	0.158
AR(2) Parameters					
Correlation $(t-1)$	0.572	0.570	0.249	0.413	0.534
Correlation $(t-2)$	0.036	0.033	0.094	0.080	-0.008

Notes: Cells report standard deviations of variance components and percentage of overall variance explained from Equation 9. Model 0 uses raw test scores, Model 1 adjusts for student demographic characteristics, Model 2 adds cubic polynomials of lagged-test scores and attendance, Model 3 restricts to students in first year in school, Model 4 uses prior-school outcomes. Models 3 and 4 not estimated for NYC because we do not observe year of first enrollment. All models include grade and year fixed effects. Demographic covariates include prior grade retention, gender, race/ethnicity, disability status, 504 plan designation, participation in migrant or Indian education program and the school averages of the preceding characteristics. All samples restricted to observations with at least 25 students in each school-by-year cell.

A Supplemental Figures and Tables

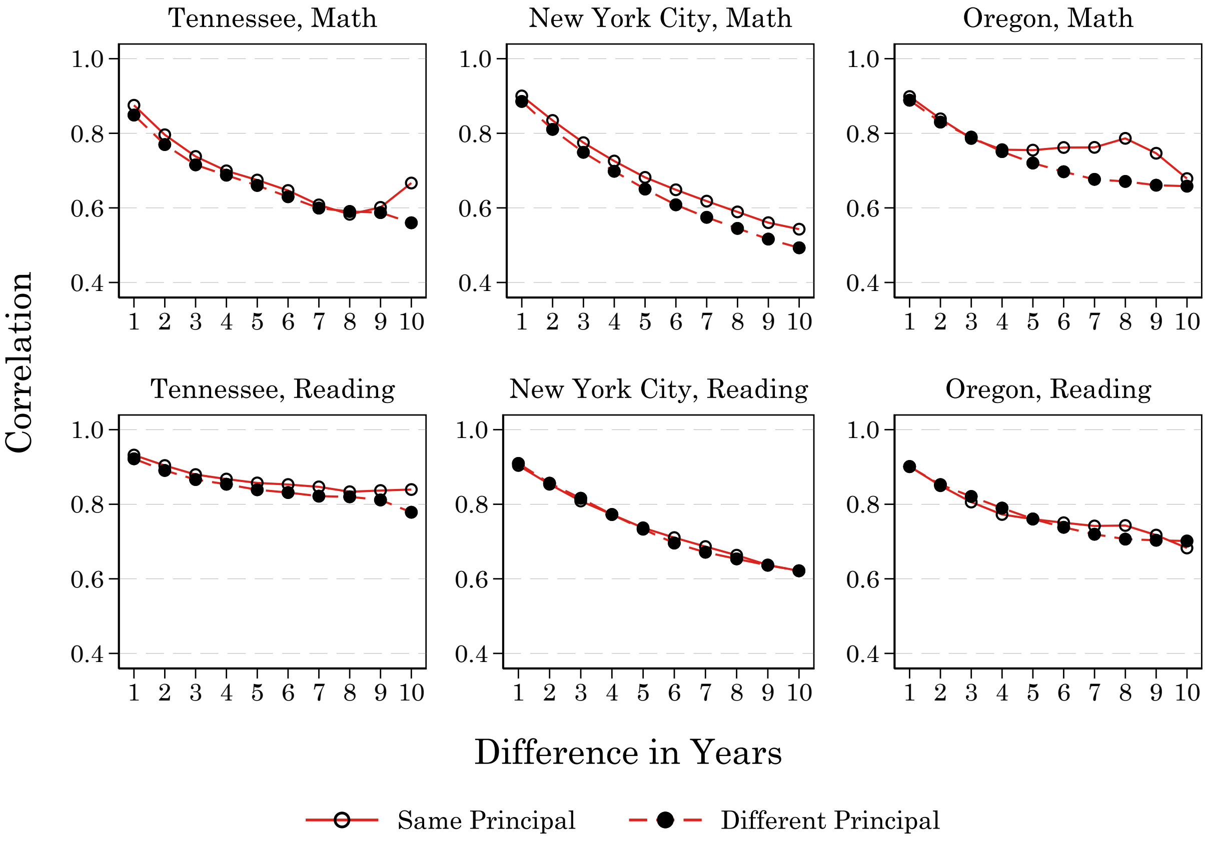


Figure A.1: Autocorrelations Within and Between Principals (Model 1)

Notes: Figures report autocorrelation "drift" vectors generated from sample-size-precision-weighted correlations of school-by-year mean residualized test score generated from Equation 5. Residualization models adjust for student demographic characteristics (Model 1).

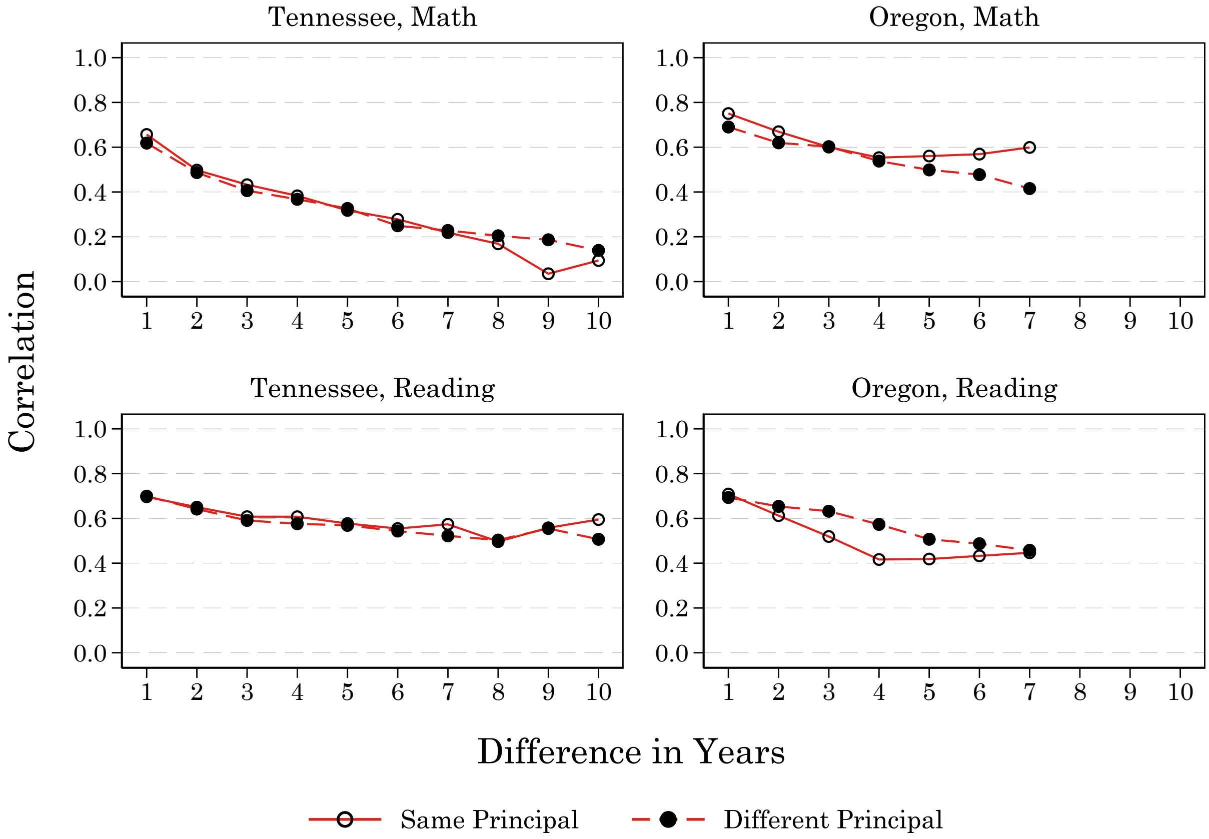


Figure A.2: Autocorrelations Within and Between Principals (Model 3)

Notes: Figures report autocorrelation "drift" vectors generated from sample-size-precision-weighted correlations of school-by-year mean residualized test score generated from Equation 5. Residualization models adjust for student demographic characteristics and cubic polynomials of lagged-test scores and attendance for new-to-school students only (Model 3). Due to very small cell sizes ($<$ 100), we do not report the correlations for 8 years of above in Oregon.

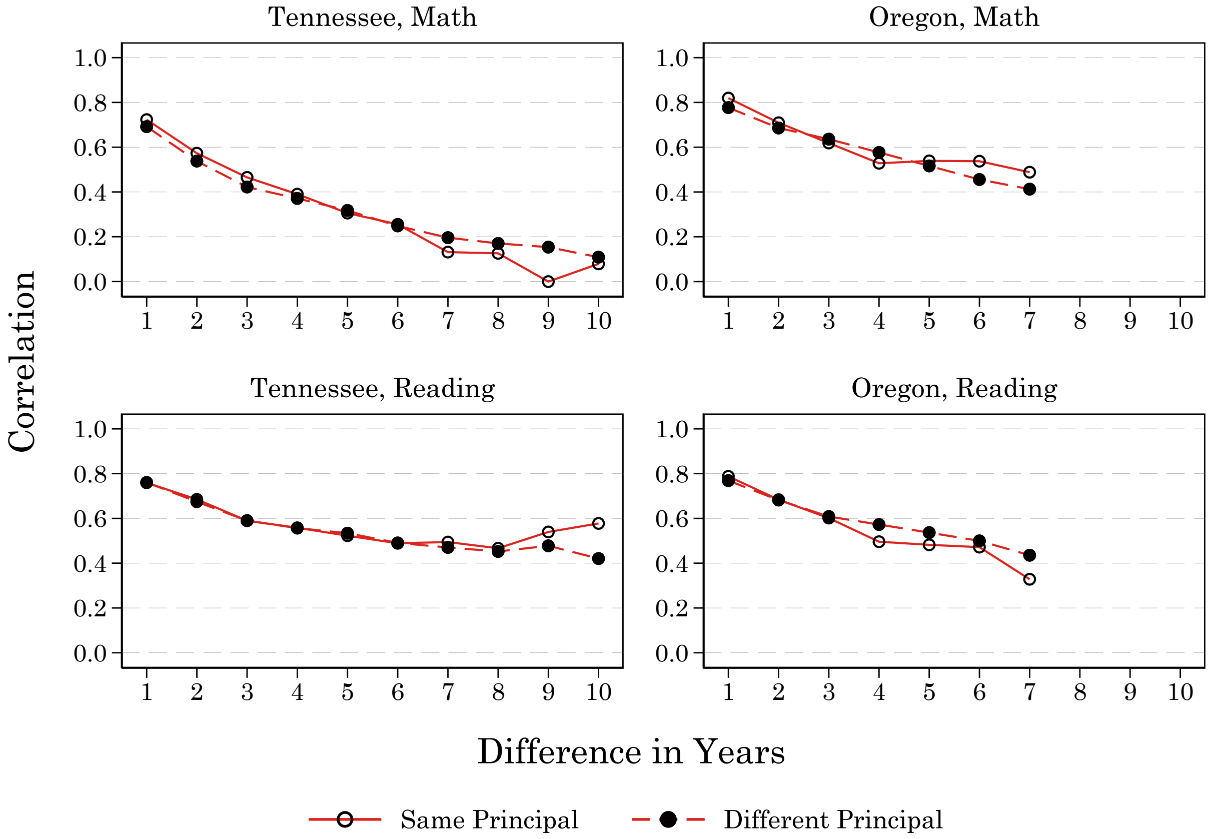
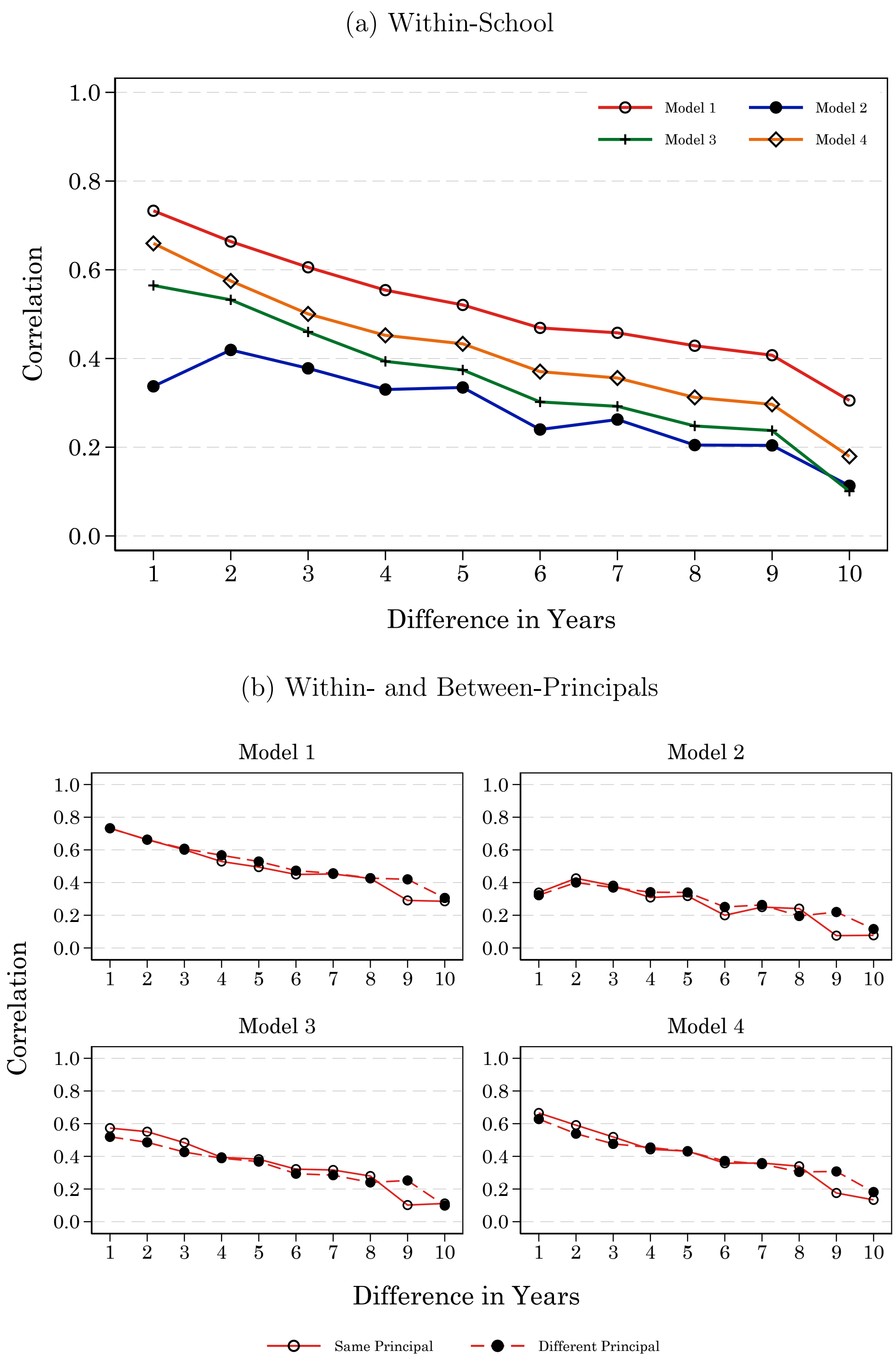


Figure A.3: Autocorrelations Within and Between Principals (Model 4)

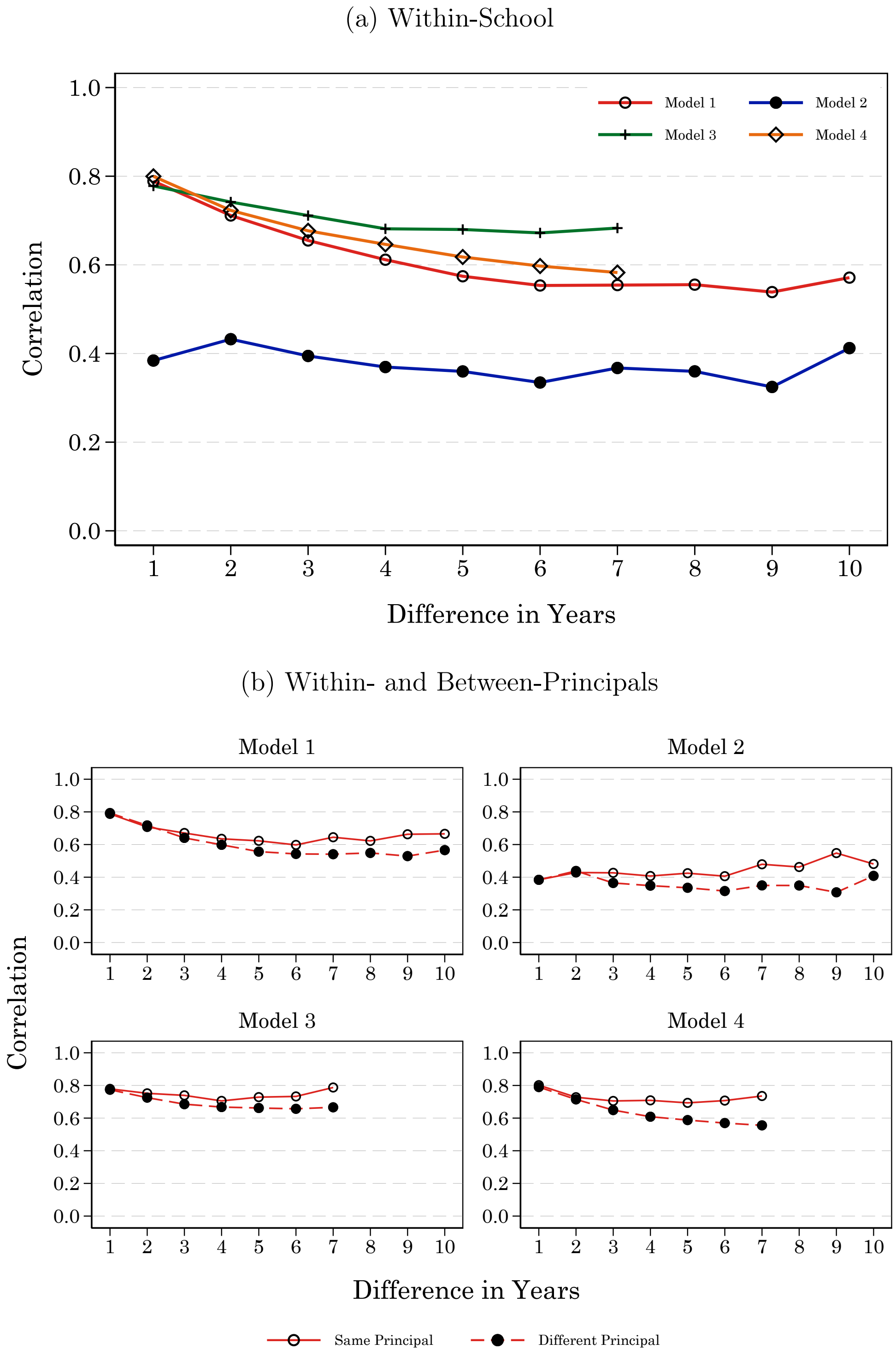
Notes: Figures report autocorrelation "drift" vectors generated from sample-size-precision-weighted correlations of school-by-year mean residualized test score generated from Equation 5. Residualization models adjust for student demographic characteristics and cubic polynomials of lagged-test scores and attendance from prior-school (Model 4). Due to very small cell sizes ($<$ 100), we do not report the correlations for 8 years of above in Oregon.


(a) Within-School

(b) Within- and Between-Principals

Figure A.4: Autocorrelation Vectors (TN Attendance)

Notes: Figures report autocorrelation "drift" vectors generated from sample-size-precision-weighted correlations of school-by-year mean residualized test score generated from Equation 5. Model 1 adjusts for student demographic characteristics, Model 2 adds cubic polynomials of lagged attendance, Model 3 restricts to students in first year in school, Model 4 uses prior-school outcomes.



(a) Within-School

(b) Within- and Between-Principals

Figure A.5: Autocorrelation Vectors (OR Attendance)

Notes: Figures report autocorrelation "drift" vectors generated from sample-size-precision-weighted correlations of school-by-year mean residualized test score generated from Equation 5. Model 1 adjusts for student demographic characteristics, Model 2 adds cubic polynomials of lagged attendance, Model 3 restricts to students in first year in school, Model 4 uses prior-school outcomes. Due to very small cell sizes (< 100), we do not report the correlations for 8 years of above in Oregon for Models 3 and 4.

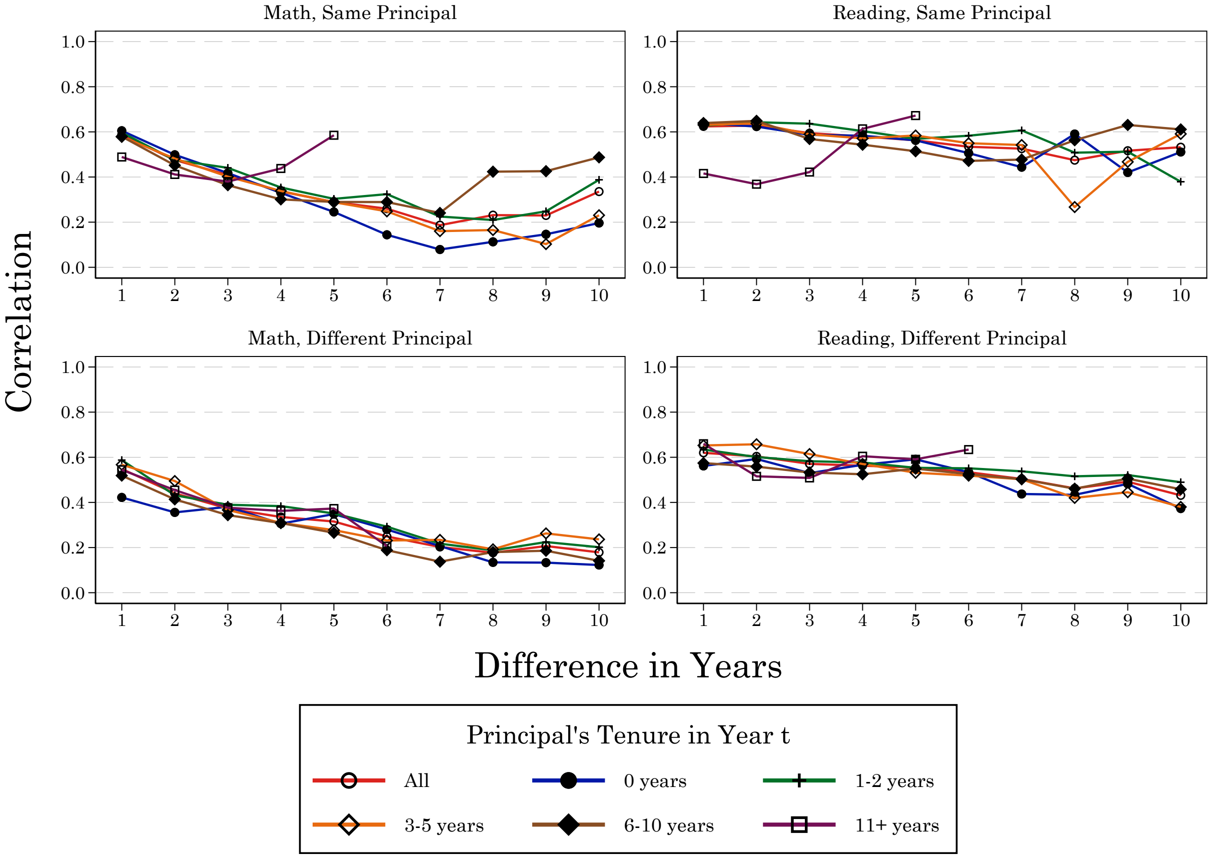


Figure A.6: Autocorrelation Vectors by Current Principal's Tenure (Tennessee)

Notes: Figures report autocorrelation "drift" vectors generated from sample-size-precision-weighted correlations of school-by-year mean residualized test score generated from Equation 5 using residualization model 2. Each plot header denotes the subject (math or reading) and whether the pair of school-by-year observations have the same principal or a different principal. In addition to a line for all school-by-year observations (i.e., the baseline results), we show lines for sub-samples defined by the years of tenure of the principal in the school in year *t*. Table A.6 shows sample sizes for each of the cells.

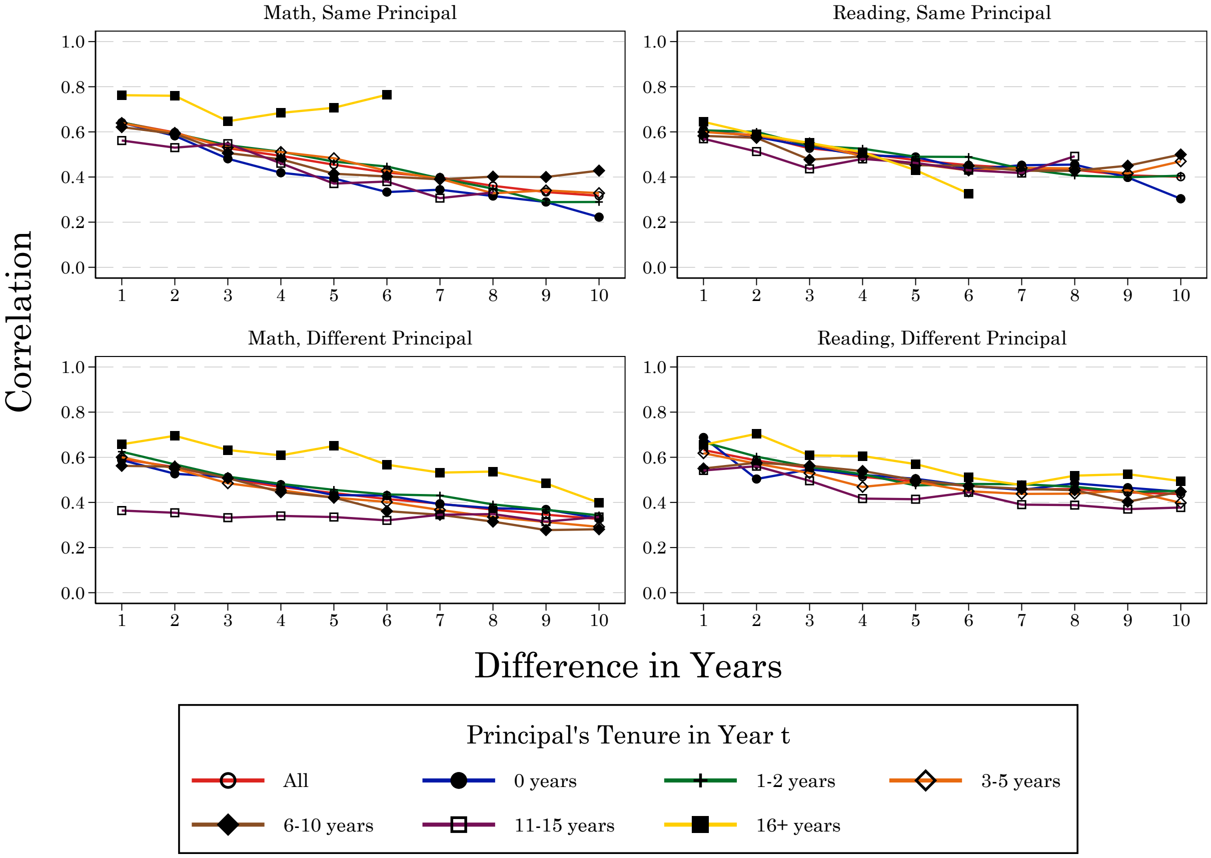


Figure A.7: Autocorrelation Vectors by Current Principal's Tenure (New York City)

Notes: Figures report autocorrelation "drift" vectors generated from sample-size-precision-weighted correlations of school-by-year mean residualized test score generated from Equation 5 using residualization model 2. Each plot header denotes the subject (math or reading) and whether the pair of school-by-year observations have the same principal or a different principal. In addition to a line for all school-by-year observations (i.e., the baseline results), we show lines for sub-samples defined by the years of tenure of the principal in the school in year t. Table A.7 shows sample sizes for each of the cells.

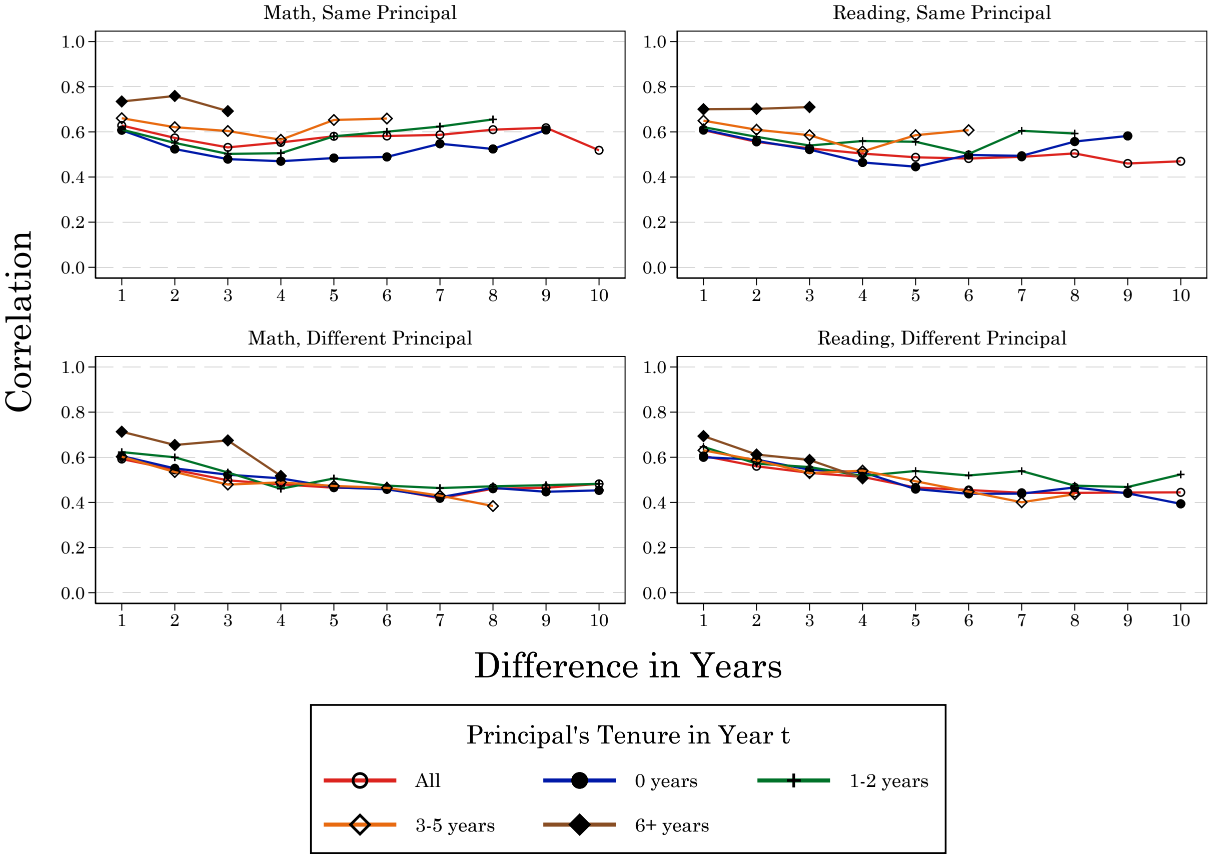


Figure A.8: Autocorrelation Vectors by Current Principal's Tenure (Oregon)

Notes: Figures report autocorrelation "drift" vectors generated from sample-size-precision-weighted correlations of school-by-year mean residualized test score generated from Equation 5 using residualization model 2. Each plot header denotes the subject (math or reading) and whether the pair of school-by-year observations have the same principal or a different principal. In addition to a line for all school-by-year observations (i.e., the baseline results), we show lines for sub-samples defined by the years of tenure of the principal in the school in year t. Table A.8 shows sample sizes for each of the cells.

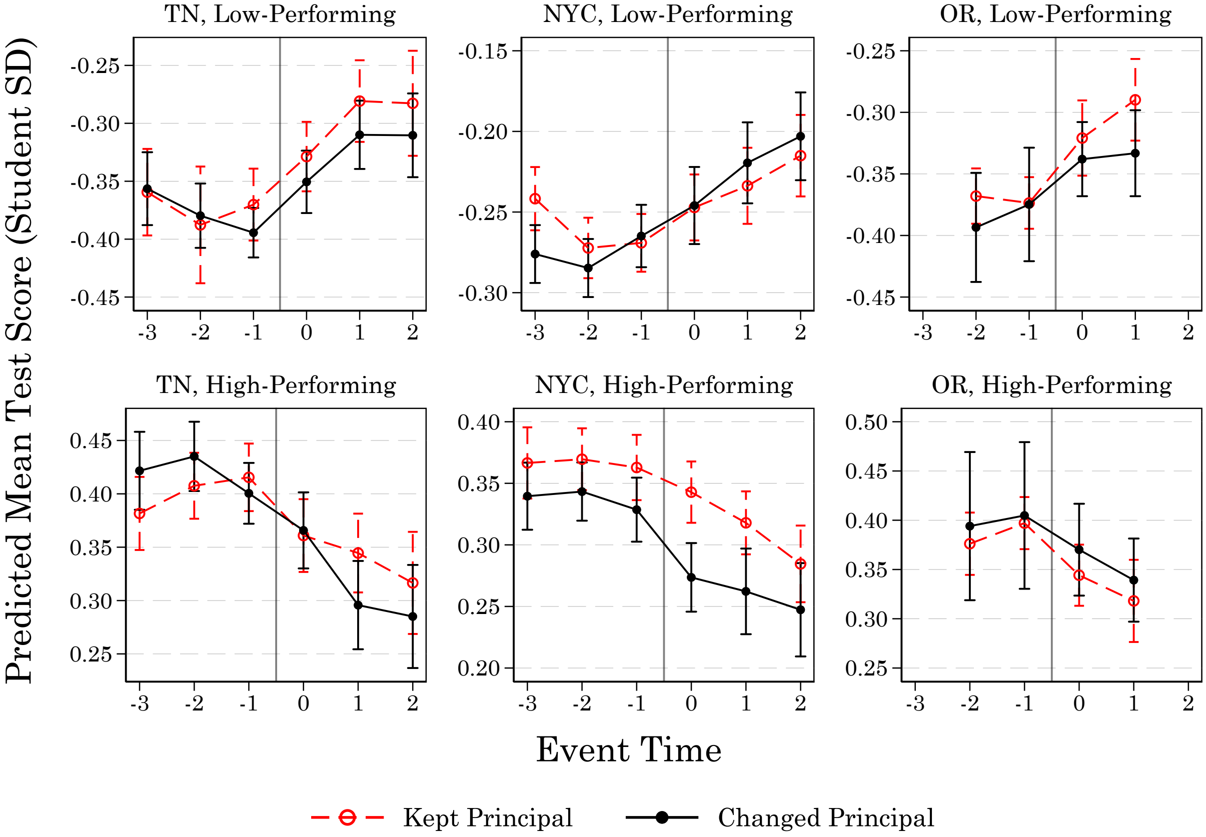


Figure A.9: Event Study (Math, Model 1)

Notes: Figures report event-study estimates and 95% confidence intervals from Equation 7. Residualization models adjust for student demographic characteristics and cubic polynomials of lagged-test scores and attendance (Model 1). High-performing and low-performing defined by top- and bottom-quartile three-year moving average of school-by-year mean test score residuals. Due to panel length limitations in Oregon, we restrict the event time to a range of four years. Standard errors adjusted for clustering at the school-spell level. School-by-year cells are weighted by the number of students contributing to the mean test score residual measure in the given year.

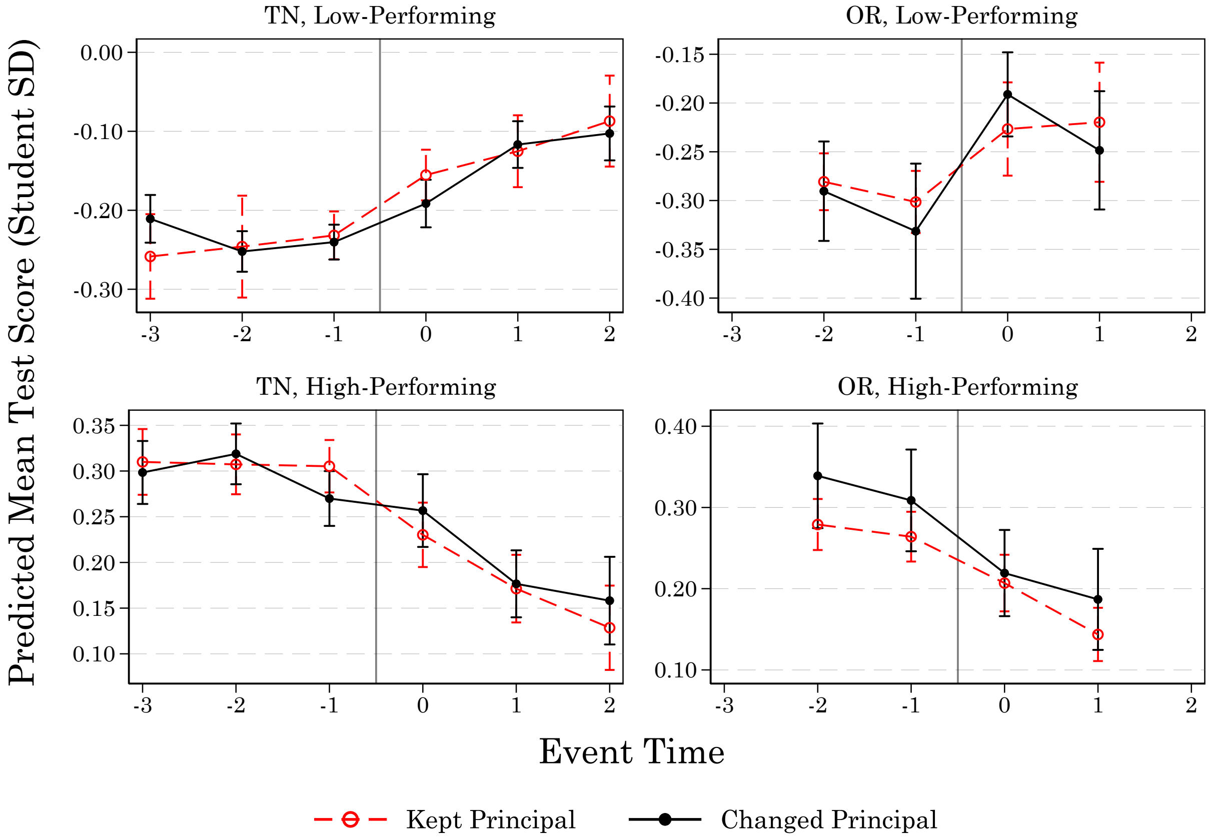


Figure A.10: Event Study (Math, Model 3)

Notes: Figures report event-study estimates and 95% confidence intervals from Equation 7. Residualization models adjust for student demographic characteristics and cubic polynomials of lagged-test scores and attendance (Model 3). High-performing and low-performing defined by top- and bottom-quartile three-year moving average of school-by-year mean test score residuals. Due to panel length limitations in Oregon, we restrict the event time to a range of four years. Standard errors adjusted for clustering at the school-spell level. School-by-year cells are weighted by the number of students contributing to the mean test score residual measure in the given year.

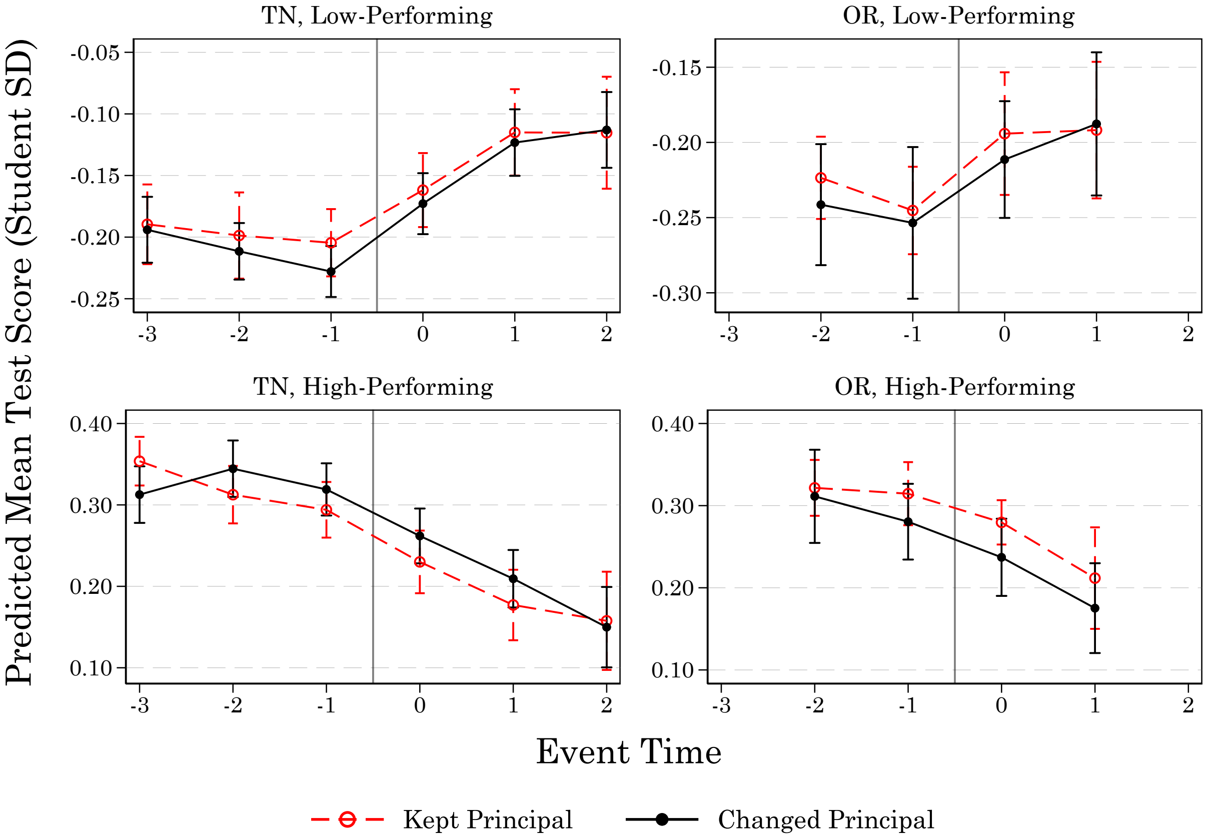


Figure A.11: Event Study (Math, Model 4)

Notes: Figures report event-study estimates and 95% confidence intervals from Equation 7. Residualization models adjust for student demographic characteristics and cubic polynomials of lagged-test scores and attendance (Model 4). High-performing and low-performing defined by top- and bottom-quartile three-year moving average of school-by-year mean test score residuals. Due to panel length limitations in Oregon, we restrict the event time to a range of four years. Standard errors adjusted for clustering at the school-spell level. School-by-year cells are weighted by the number of students contributing to the mean test score residual measure in the given year.

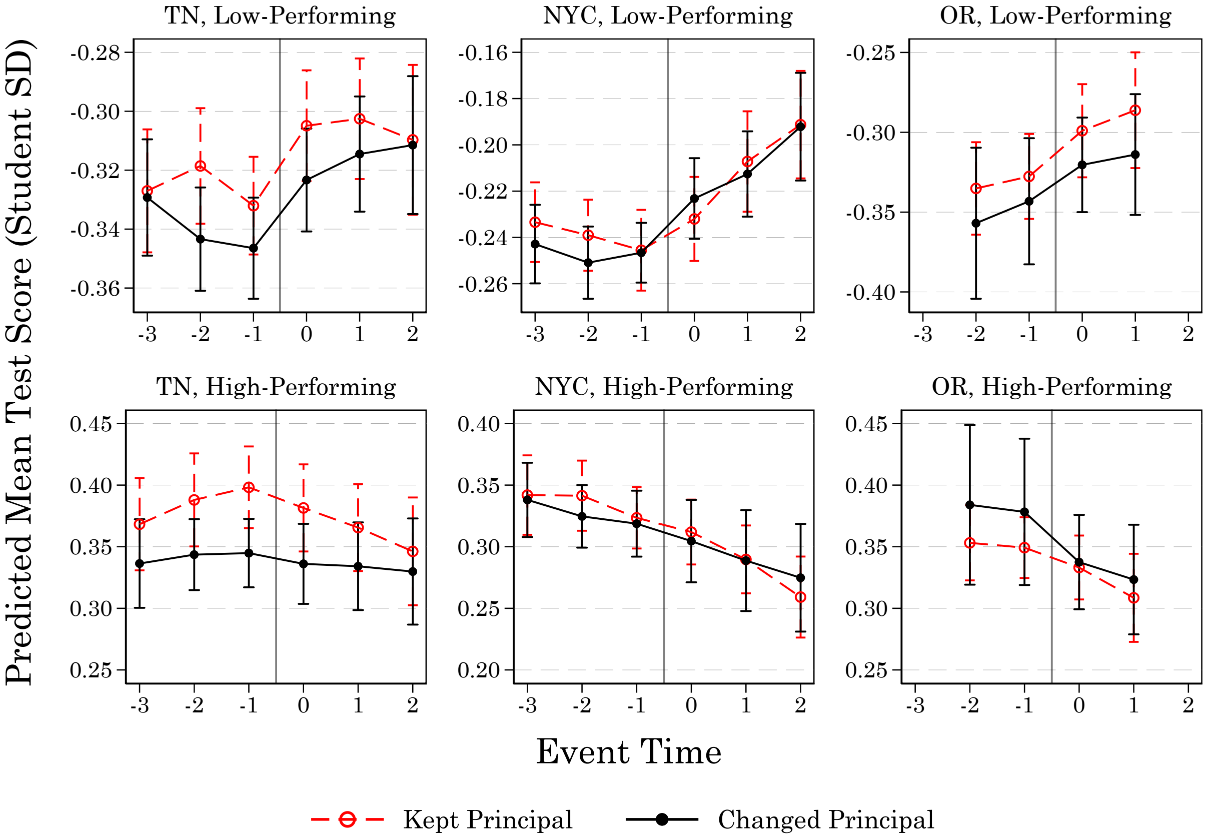


Figure A.12: Event Study (Read, Model 1)

Notes: Figures report event-study estimates and 95% confidence intervals from Equation 7. Residualization models adjust for student demographic characteristics and cubic polynomials of lagged-test scores and attendance (Model 1). High-performing and low-performing defined by top- and bottom-quartile three-year moving average of school-by-year mean test score residuals. Due to panel length limitations in Oregon, we restrict the event time to a range of four years. Standard errors adjusted for clustering at the school-spell level. School-by-year cells are weighted by the number of students contributing to the mean test score residual measure in the given year.

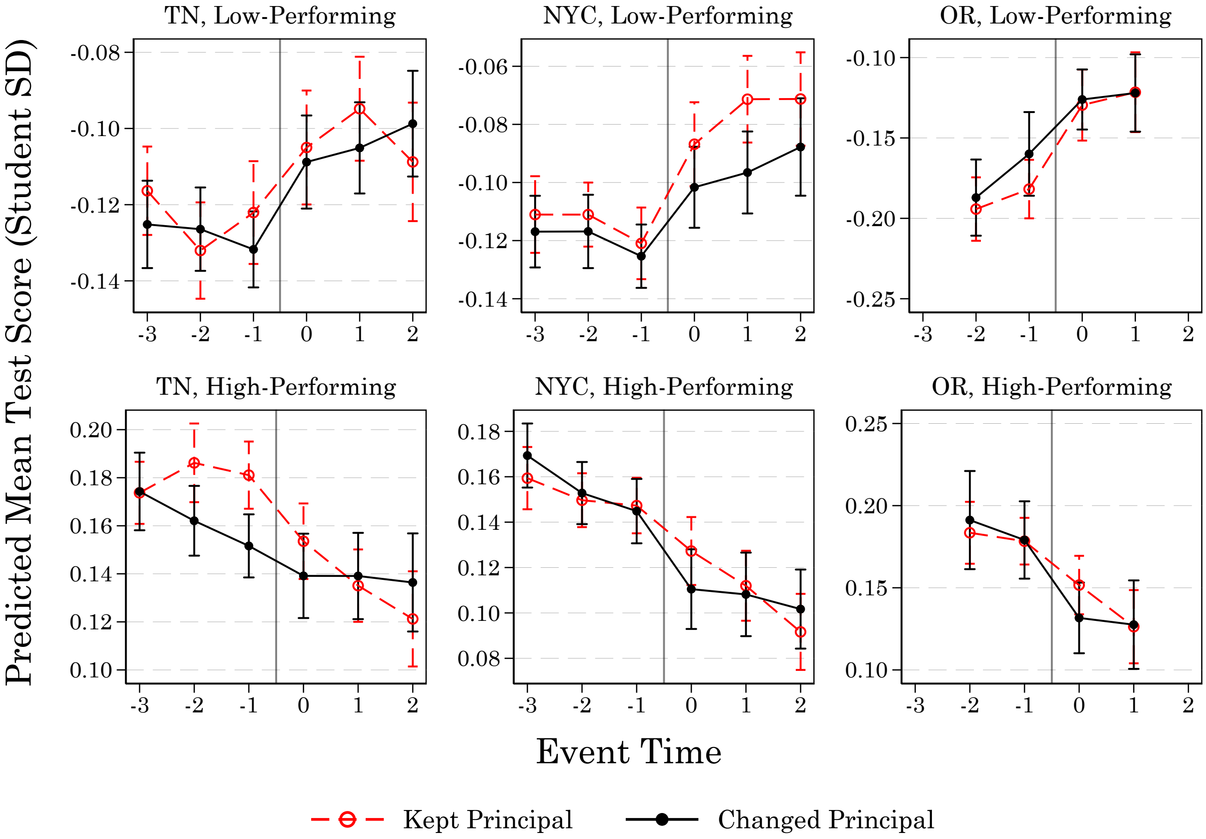


Figure A.13: Event Study (Read, Model 2)

Notes: Figures report event-study estimates and 95% confidence intervals from Equation 7. Residualization models adjust for student demographic characteristics and cubic polynomials of lagged-test scores and attendance (Model 2). High-performing and low-performing defined by top- and bottom-quartile three-year moving average of school-by-year mean test score residuals. Due to panel length limitations in Oregon, we restrict the event time to a range of four years. Standard errors adjusted for clustering at the school-spell level. School-by-year cells are weighted by the number of students contributing to the mean test score residual measure in the given year.

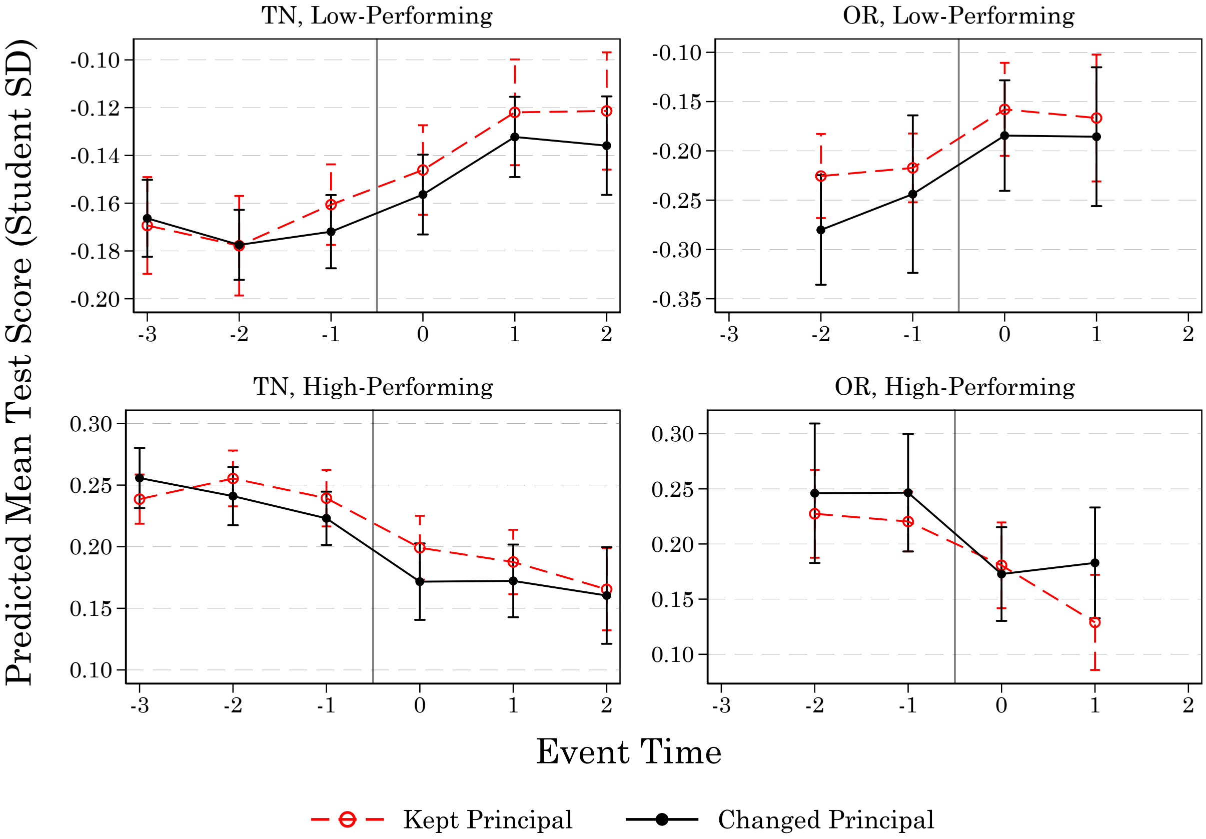


Figure A.14: Event Study (Read, Model 3)

Notes: Figures report event-study estimates and 95% confidence intervals from Equation 7. Residualization models adjust for student demographic characteristics and cubic polynomials of lagged-test scores and attendance (Model 3). High-performing and low-performing defined by top- and bottom-quartile three-year moving average of school-by-year mean test score residuals. Due to panel length limitations in Oregon, we restrict the event time to a range of four years. Standard errors adjusted for clustering at the school-spell level. School-by-year cells are weighted by the number of students contributing to the mean test score residual measure in the given year.

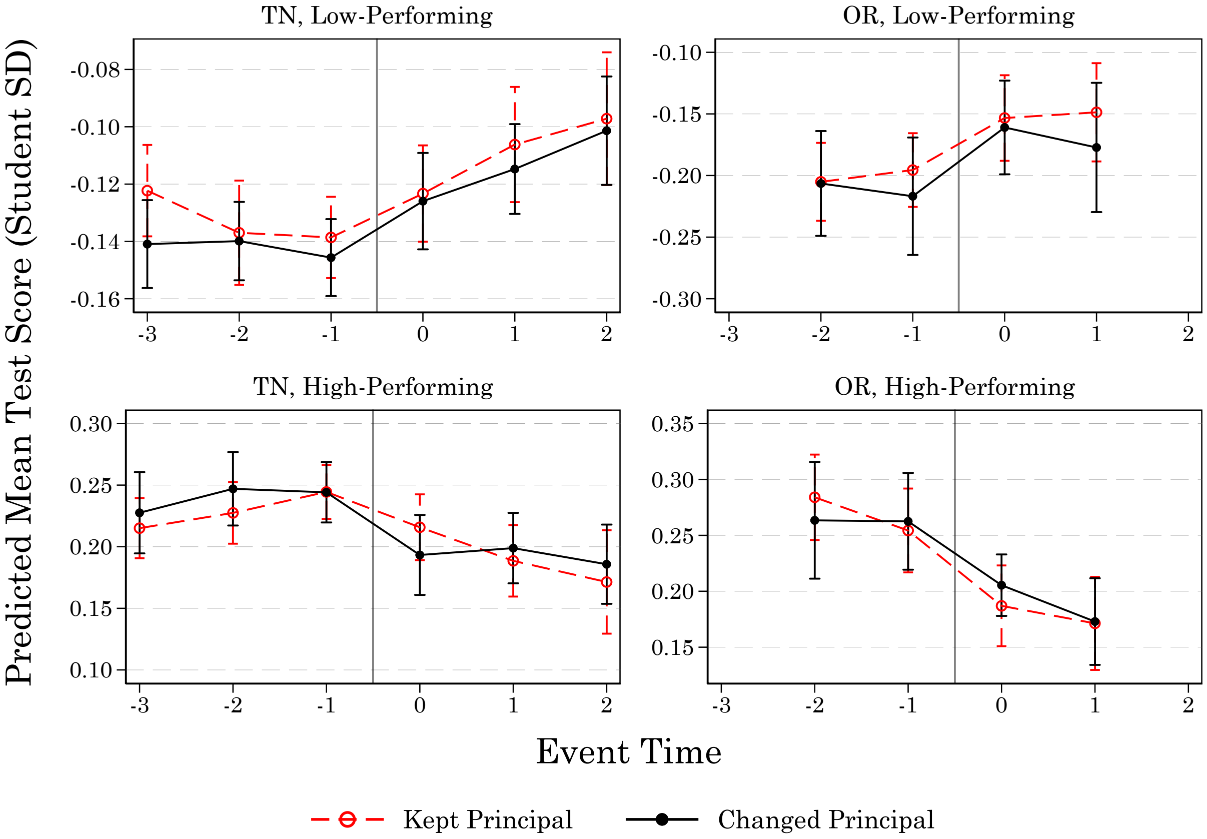


Figure A.15: Event Study (Read, Model 4)

Notes: Figures report event-study estimates and 95% confidence intervals from Equation 7. Residualization models adjust for student demographic characteristics and cubic polynomials of lagged-test scores and attendance (Model 4). High-performing and low-performing defined by top- and bottom-quartile three-year moving average of school-by-year mean test score residuals. Due to panel length limitations in Oregon, we restrict the event time to a range of four years. Standard errors adjusted for clustering at the school-spell level. School-by-year cells are weighted by the number of students contributing to the mean test score residual measure in the given year.

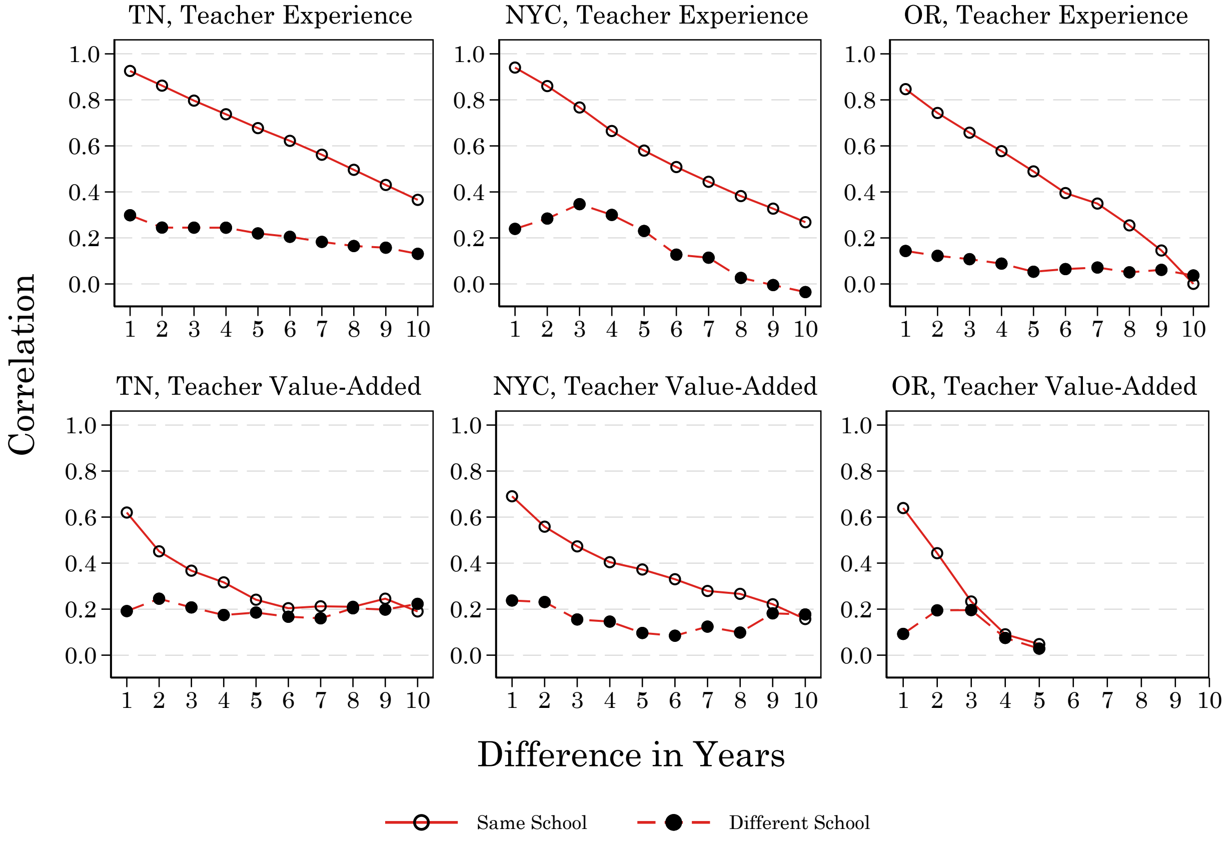


Figure A.16: Principal Autocorrelation Vectors for Teacher Composition

Notes: Figures report autocorrelation "drift" vectors generated from sample-size-precision-weighted correlations of principal-by-year mean teacher experience and value-added (pooling math and reading teachers). Correlations are between year t and $t+x$ for the same principal, where x is denoted by the x-axis value. Same school denotes the sub-sample of principal-by-year pairs where the school is the same in both years. Different school denotes the sub-sample where the school in year t is different than year $t+x$. For teacher experience, principal-by-year cells are weighted by the number of teachers in the school. For VA, principal-by-year cells are weighted by the number of teachers with a VA estimate.

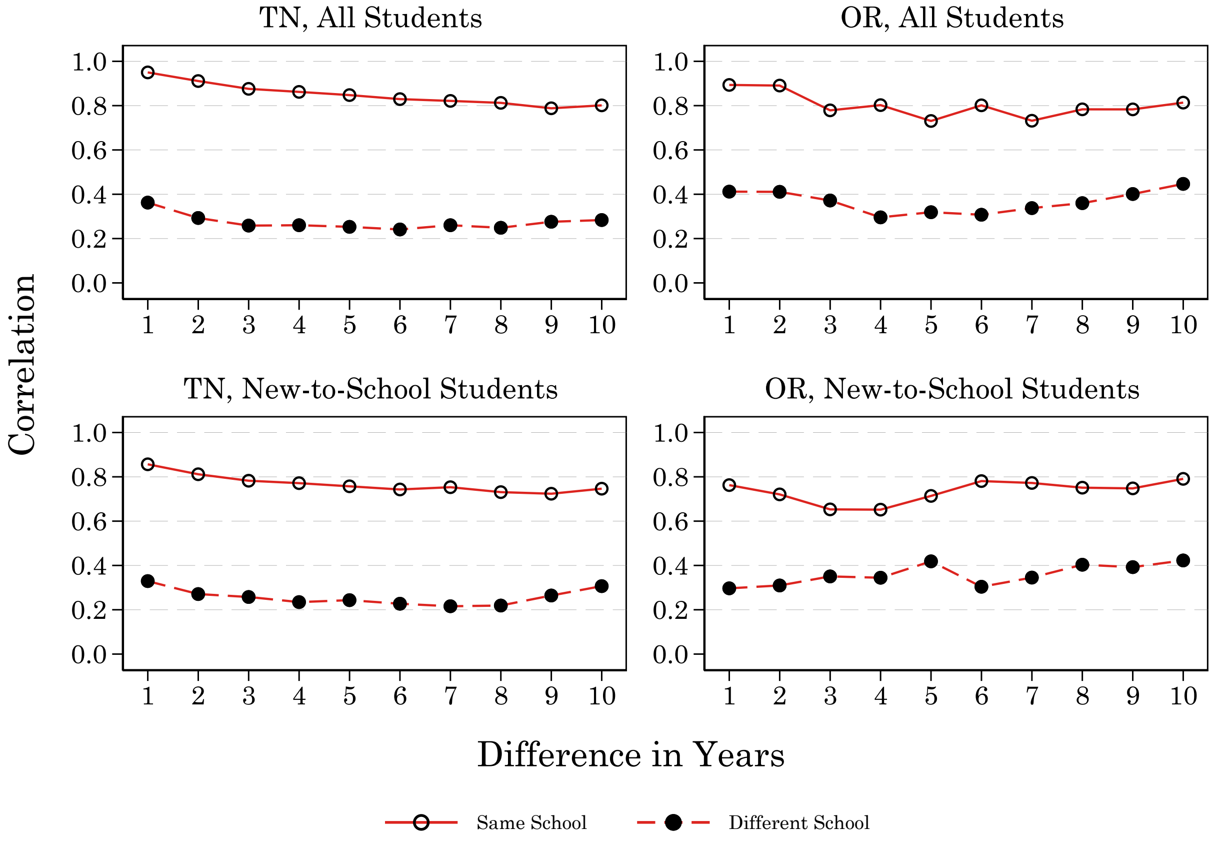


Figure A.17: Principal Autocorrelation Vectors for Student Composition

Notes: Figures report autocorrelation "drift" vectors generated from sample-size-precision-weighted correlations of principal-by-year prior-school outcomes for both new-to-school and all students generated from Equation 5. Correlations are between year t and $t+x$ for the same principal, where x is denoted by the x-axis value. Same school denotes the sub-sample of principal-by-year pairs where the school is the same in both years. Different school denotes the sub-sample where the school in year t is different than year $t+x$. OR results exclude HS students as very few new-to-school students have a prior-school *and* a current-year score because only a small number of 9^{th}-grade students appear in our sample prior to 2014 and none afterwards.

Table A.1: Variance Decomposition (Reading)

	Model 0	Model 1	Model 2	Model 3	Model 4
Panel A: Tennessee					
Random Effects Parameters (SD)					
School	0.430	0.288	0.114	0.161	0.139
Principal	0.092	0.084	0.044	0.055	0.064
Residual	0.119	0.113	0.099	0.113	0.096
Variance Components (%)					
School	89.2	80.7	52.6	62.1	59.2
Principal	4.0	6.9	8.0	7.3	12.7
Residual	6.8	12.4	39.4	30.7	28.2
N (Schools)	1841	1841	1841	1418	1304
N (Principal-by-School)	4796	4796	4796	3264	3012
N (School-by-Year Cells)	17577	17577	17577	10488	9942
Mean Students Per Cell	326	326	326	182	364
Panel B: New York City					
Random Effects Parameters (SD)					
School	0.486	0.226	0.110		
Principal	0.143	0.098	0.040		
Residual	0.135	0.122	0.101		
Variance Components (%)					
School	86.0	67.5	50.6		
Principal	7.4	12.7	6.7		
Residual	6.6	19.8	42.7		
N (Schools)	1316	1316	1316		
N (Principal-by-School)	3490	3490	3490		
N (School-by-Year Cells)	18337	18337	18337		
Mean Students Per Cell	326	326	326		
Panel C: Oregon					
Random Effects Parameters (SD)					
School	0.340	0.274	0.131	0.168	0.163
Principal	0.097	0.090	0.059	0.075	0.073
Residual	0.126	0.121	0.124	0.129	0.110
Variance Components (%)					
School	82.0	76.9	47.7	55.8	60.2
Principal	6.6	8.2	9.6	11.2	12.0
Residual	11.4	14.9	42.7	32.9	27.7
N (Schools)	1271	1271	1271	841	820
N (Principal-by-School)	3562	3562	3562	1817	1657
N (School-by-Year Cells)	11819	11819	11819	4864	4326
Mean Students Per Cell	239	239	239	144	229
Student Characteristics		✓	✓	✓	✓
Prior-Year Test Scores			✓	✓	
New-to-School Students Only				✓	
Prior-School Test Scores					✓

Notes: Cells report standard deviations of variance components and percentage of overall variance explained from Equation 6. Model 0 uses raw test scores, Model 1 adjusts for student demographic characteristics, Model 2 adds cubic polynomials of lagged-test scores and attendance, Model 3 restricts to students in first year in school, Model 4 uses prior-school outcomes. Models 3 and 4 not estimated for NYC because we do not observe year of first enrollment. All models include grade and year fixed effects. Demographic covariates include prior grade retention, gender, race/ethnicity, disability status, 504 plan designation, participation in migrant or Indian education program and the school averages of the preceding characteristics. All samples restricted to observations with at least 25 students in each school-by-year cell.

Table A.2: Variance Decomposition (Attendance Rate, TN)

	Model 0	Model 1	Model 2	Model 3	Model 4
Panel A: Baseline					
School (SD)	0.257	0.258	0.161	0.191	0.196
%	58.0	58.5	38.6	35.0	41.4
Principal (SD)	0.126	0.124	0.070	0.112	0.122
%	14.0	13.5	7.3	12.1	16.0
Residual (SD)	0.179	0.178	0.191	0.235	0.199
%	28.1	28.0	54.1	52.9	42.5
N (Schools)	1938	1938	1938	1906	1918
N (Principals)	5095	5095	5095	4927	4963
N (School-by-Year Cells)	20153	20153	20153	19011	19435
Mean Students Per Cell	569	569	569	180	352
Panel B: AR(1) Error Structure					
School (SD)	0.256	0.256	0.162	0.191	0.195
%	57.6	58.3	38.6	35.1	41.2
Principal (SD)	0.072	0.070	0.077	0.093	0.078
%	4.5	4.4	8.9	8.3	6.6
Residual (SD)	0.207	0.205	0.188	0.243	0.219
%	37.8	37.3	52.5	56.7	52.2
AR(1) Parameters					
Correlation $(t-1)$	0.445	0.434	-0.083	0.152	0.346
Panel C: AR(2) Error Structure					
School (SD)	0.254	0.255	0.161	0.191	0.194
%	56.8	57.5	38.5	34.9	40.9
Principal (SD)	0.000	0.000	0.066	0.041	0.000
%	0.0	0.0	6.4	1.6	0.0
Residual (SD)	0.221	0.219	0.193	0.257	0.233
%	43.2	42.5	55.1	63.4	59.1
AR(2) Parameters					
Correlation $(t-1)$	0.437	0.431	-0.033	0.203	0.370
Correlation $(t-2)$	0.149	0.142	0.112	0.158	0.122
Student Characteristics		✓	✓	✓	✓
Prior-Year Attendance			✓	✓	
New-to-School Students Only				✓	
Prior-School Attendance					✓

Notes: Cells report standard deviations of variance components and percentage of overall variance explained from Equation 9. Model 0 uses raw test scores, Model 1 adjusts for student demographic characteristics, Model 2 adds cubic polynomials of lagged attendance, Model 3 restricts to students in first year in school, Model 4 uses prior-school outcomes. All models include grade and year fixed effects. Demographic covariates include prior grade retention, gender, race/ethnicity, disability status, 504 plan designation, participation in migrant or Indian education program and the school averages of the preceding characteristics. All samples restricted to observations with at least 25 students in each school-by-year cell.

Table A.3: Variance Decomposition (Attendance Rate, Oregon)

	Model 0	Model 1	Model 2	Model 3	Model 4
Panel A: Baseline					
School (SD)	0.219	0.212	0.122	0.260	0.252
%	69.0	67.6	47.1	69.4	69.2
Principal (SD)	0.082	0.082	0.035	0.071	0.087
%	9.8	10.1	3.8	5.1	8.3
Residual (SD)	0.121	0.122	0.124	0.158	0.144
%	21.3	22.3	49.0	25.5	22.5
N (Schools)	1347	1347	1347	1226	863
N (Principals)	3767	3767	3767	3341	1885
N (School-by-Year Cells)	12449	12449	12449	10950	5296
Mean Students Per Cell	435	435	435	132	240
Panel B: AR(1) Error Structure					
School (SD)	0.219	0.212	0.122	0.260	0.254
%	69.4	68.0	47.0	69.5	70.1
Principal (SD)	0.028	0.025	0.042	0.054	0.035
%	1.2	0.9	5.5	3.0	1.3
Residual (SD)	0.143	0.143	0.122	0.164	0.162
%	29.5	31.1	47.5	27.4	28.6
AR(1) Parameters					
Correlation $(t-1)$	0.452	0.457	-0.078	0.148	0.351
Panel C: AR(2) Error Structure					
School (SD)	0.218	0.212	0.122	0.261	0.254
%	69.0	67.7	47.1	69.6	70.1
Principal (SD)	0.000	0.000	0.034	0.033	0.031
%	0.0	0.0	3.7	1.1	1.1
Residual (SD)	0.146	0.146	0.125	0.169	0.163
%	31.0	32.3	49.2	29.3	28.9
AR(2) Parameters					
Correlation $(t-1)$	0.436	0.440	-0.037	0.183	0.353
Correlation $(t-2)$	0.087	0.079	0.086	0.094	0.010
Student Characteristics		✓	✓	✓	✓
Prior-Year Attendance			✓	✓	
New-to-School Students Only				✓	
Prior-School Attendance					✓

Notes: Cells report standard deviations of variance components and percentage of overall variance explained from Equation 9. Model 0 uses raw test scores, Model 1 adjusts for student demographic characteristics, Model 2 adds cubic polynomials of lagged attendance, Model 3 restricts to students in first year in school, Model 4 uses prior-school outcomes. All models include grade and year fixed effects. Demographic covariates include prior grade retention, gender, race/ethnicity, disability status, 504 plan designation, participation in migrant or Indian education program and the school averages of the preceding characteristics. All samples restricted to observations with at least 25 students in each school-by-year cell.

Table A.4: Sample Sizes for Figures 1 and 2 (Residualization Model 2)

	Tennessee			New York City			Oregon		
	All	Same	Diff	All	Same	Diff	All	Same	Diff
Panel A: Math									
Diff = 1 year	14306	11846	2460	16863	14848	2015	10070	8066	2004
Diff = 2 years	12452	8429	4023	15593	11900	3693	8917	5543	3374
Diff = 3 years	10966	6038	4928	14361	9437	4924	7855	3675	4180
Diff = 4 years	10523	4670	5853	13190	7462	5728	6880	2374	4506
Diff = 5 years	8939	3181	5758	12033	5897	6136	5945	1470	4475
Diff = 6 years	7379	2125	5254	10907	4629	6278	5056	894	4162
Diff = 7 years	5857	1312	4545	9798	3597	6201	4212	534	3678
Diff = 8 years	4375	786	3589	8719	2745	5974	3343	303	3040
Diff = 9 years	4015	599	3416	7675	2019	5656	2491	171	2320
Diff = 10 years	2766	340	2426	6661	1428	5233	1645	86	1559
Panel B: Reading									
Diff = 1 year	14348	11880	2468	16849	14832	2017	10071	8065	2006
Diff = 2 years	12495	8457	4038	15576	11883	3693	8919	5544	3375
Diff = 3 years	11003	6056	4947	14348	9422	4926	7855	3676	4179
Diff = 4 years	10560	4685	5875	13178	7450	5728	6877	2374	4503
Diff = 5 years	8964	3189	5775	12020	5888	6132	5942	1469	4473
Diff = 6 years	7403	2131	5272	10895	4623	6272	5054	893	4161
Diff = 7 years	5878	1313	4565	9786	3591	6195	4217	535	3682
Diff = 8 years	4384	785	3599	8707	2741	5966	3344	303	3041
Diff = 9 years	4018	598	3420	7666	2016	5650	2491	171	2320
Diff = 10 years	2765	339	2426	6652	1426	5226	1646	86	1560

Table A.5: Replication of Branch, Hanushek, and Rivkin (2012)

	Tennessee				New York City				Oregon			
	Math		Read		Math		Read		Math		Read	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Different Principal	0.011***	0.005***	0.002***	0.000	0.004***	0.000	0.001***	-0.001***	0.009***	0.003***	0.008***	0.002**
	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)
Diff = 1 year (base)												
Diff = 2 years		0.008***		0.000		0.002***		0.001		0.003**		0.002**
		(0.001)		(0.000)		(0.000)		(0.000)		(0.001)		(0.001)
Diff = 3 years		0.011***		0.001**		0.004***		0.002***		0.007***		0.004***
		(0.001)		(0.000)		(0.000)		(0.000)		(0.001)		(0.001)
Diff = 4 years		0.014***		0.001***		0.006***		0.003***		0.008***		0.006***
		(0.001)		(0.000)		(0.000)		(0.000)		(0.001)		(0.001)
Diff = 5 years		0.015***		0.002***		0.007***		0.004***		0.008***		0.008***
		(0.001)		(0.000)		(0.000)		(0.000)		(0.001)		(0.001)
Diff = 6 years		0.018***		0.003***		0.008***		0.005***		0.010***		0.010***
		(0.001)		(0.000)		(0.000)		(0.000)		(0.001)		(0.001)
Diff = 7 years		0.021***		0.004***		0.009***		0.005***		0.014***		0.012***
		(0.001)		(0.001)		(0.000)		(0.000)		(0.001)		(0.001)
Diff = 8 years		0.021***		0.005***		0.009***		0.005***		0.013***		0.014***
		(0.001)		(0.001)		(0.001)		(0.000)		(0.001)		(0.001)
Diff = 9 years		0.017***		0.004***		0.010***		0.005***		0.016***		0.017***
		(0.001)		(0.001)		(0.001)		(0.000)		(0.002)		(0.001)
Diff = 10 years		0.013***		0.005***		0.010***		0.005***		0.024***		0.026***
		(0.002)		(0.001)		(0.001)		(0.000)		(0.002)		(0.002)
Diff = 11 years		0.011***		0.003***		0.010***		0.005***		0.045***		0.047***
		(0.002)		(0.001)		(0.001)		(0.000)		(0.003)		(0.002)
Diff = 12 years						0.010***		0.006***				
						(0.001)		(0.001)				
Diff = 13 years						0.009***		0.006***				
						(0.001)		(0.001)				
Diff = 14 years						0.009***		0.007***				
						(0.001)		(0.001)				
Diff = 15 years						0.009***		0.005***				
						(0.001)		(0.001)				
Diff = 16 years						0.010***		0.005***				
						(0.001)		(0.001)				
Diff = 17 years						0.008***		0.005***				
						(0.001)		(0.001)				
Constant	0.035***	0.026***	0.013***	0.012***	0.018***	0.014***	0.014***	0.012***	0.028***	0.024***	0.022***	0.019***
	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.001)
Within-school Variance	0.005	0.002	0.001	0.000	0.002	0.000	0.001	0.000	0.005	0.002	0.004	0.001
Within-school SD	0.072	0.049	0.028	0.012	0.042	0.010	0.026	0.000	0.068	0.040	0.062	0.028
N	82954	82954	83193	83193	137470	137470	137318	137318	57228	57228	57231	57231

Notes: Coefficients are from models estimated via OLS predicting the squared difference between school-by-year mean residuals in year t and t^* as a function of whether the principal is different in those years. Even columns add controls for the difference in time between the two school-by-year cells that form the dependent variable. The within-school variance is equal to one-half of the coefficient on different principal, per the framework in Branch, Hanushek, and Rivkin (2012). Standard errors shown in parentheses. N refers to the total number of pairs of school-by-year cells. Each pair is weighted by the sum of the number of students that contribute to the school-by-year mean residual test score.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A.6: Sample Sizes for Figure A.6

		Principal's Tenure in Year t				
	All	0	1–2	3–5	6–10	11+
Panel A: Same Principal						
Diff = 1 year	11846	2146	3414	3042	2762	482
Diff = 2 years	8429	1564	2458	2129	2006	272
Diff = 3 years	6038	1140	1762	1518	1460	158
Diff = 4 years	4670	900	1377	1152	1073	168
Diff = 5 years	3181	604	953	766	766	92
Diff = 6 years	2125	443	609	495	547	31
Diff = 7 years	1312	255	368	327	362	0
Diff = 8 years	786	156	205	214	211	0
Diff = 9 years	599	97	169	168	165	0
Diff = 10 years	340	61	97	91	91	0
Panel B: Different Principal						
Diff = 1 year	2460	267	688	721	658	126
Diff = 2 years	4023	509	1154	1124	1075	161
Diff = 3 years	4928	695	1447	1304	1341	141
Diff = 4 years	5853	894	1693	1517	1520	229
Diff = 5 years	5758	934	1612	1515	1504	193
Diff = 6 years	5254	843	1446	1410	1445	110
Diff = 7 years	4545	708	1284	1196	1357	0
Diff = 8 years	3589	573	1045	885	1086	0
Diff = 9 years	3416	580	944	854	1038	0
Diff = 10 years	2426	429	638	630	729	0

Table A.7: Sample Sizes for Figure A.7

		Principal's Tenure in Year t					
	All	0	1–2	3–5	6–10	11–15	16+
Panel A: Same Principal							
Diff = 1 year	15145	2328	3974	4009	3552	994	288
Diff = 2 years	12150	1991	3234	3278	2788	660	199
Diff = 3 years	9641	1619	2645	2688	2115	432	142
Diff = 4 years	7629	1319	2173	2215	1538	284	100
Diff = 5 years	6034	1086	1805	1799	1068	203	73
Diff = 6 years	4742	901	1509	1421	713	144	54
Diff = 7 years	3689	756	1260	1071	467	95	40
Diff = 8 years	2818	641	1014	773	295	63	32
Diff = 9 years	2079	528	776	519	189	42	25
Diff = 10 years	1472	405	570	320	130	27	20
Panel B: Different Principal							
Diff = 1 year	2056	273	407	546	532	216	82
Diff = 2 years	3755	426	848	987	980	372	142
Diff = 3 years	5013	617	1186	1295	1276	465	174
Diff = 4 years	5830	764	1403	1530	1439	501	193
Diff = 5 years	6253	861	1528	1690	1474	500	200
Diff = 6 years	6394	902	1609	1780	1417	486	200
Diff = 7 years	6315	917	1647	1788	1304	464	195
Diff = 8 years	6082	922	1679	1671	1180	449	181
Diff = 9 years	5755	929	1654	1496	1069	440	167
Diff = 10 years	5325	928	1565	1301	951	428	152

Table A.8: Sample Sizes for Figure A.8

		Principal's Tenure in Year t			
	All	0	1–2	3–5	6+
Panel A: Same Principal					
Diff = 1 year	8066	1975	2461	1436	379
Diff = 2 years	5543	1465	1706	901	197
Diff = 3 years	3675	1036	1127	542	103
Diff = 4 years	2374	715	717	318	41
Diff = 5 years	1470	452	456	175	9
Diff = 6 years	894	291	264	96	0
Diff = 7 years	534	185	152	39	0
Diff = 8 years	303	100	93	8	0
Diff = 9 years	171	66	38	0	0
Diff = 10 years	86	33	8	0	0
Panel B: Different Principal					
Diff = 1 year	2004	336	534	417	133
Diff = 2 years	3374	598	894	673	182
Diff = 3 years	4180	786	1118	748	166
Diff = 4 years	4506	873	1199	735	113
Diff = 5 years	4475	932	1151	648	46
Diff = 6 years	4162	878	1040	498	0
Diff = 7 years	3678	785	876	305	0
Diff = 8 years	3040	676	668	120	0
Diff = 9 years	2320	526	447	0	0
Diff = 10 years	1559	390	169	0	0

Table A.9: 2×2 Difference-in-Differences Estimates (TN)

	Math				Reading			
	Model 1	Model 2	Model 3	Model 4	Model 1	Model 2	Model 3	Model 4
Panel A: Low-Performing Schools								
Post	0.071***	0.069***	0.112***	0.064***	0.023**	0.021***	0.039***	0.026**
	(0.017)	(0.012)	(0.026)	(0.015)	(0.009)	(0.005)	(0.009)	(0.009)
Changed Principal	0.000	0.001	0.014	-0.007	-0.009	-0.001	-0.001	-0.004
	(0.018)	(0.012)	(0.023)	(0.013)	(0.010)	(0.004)	(0.008)	(0.007)
Post x Changed Principal	-0.022	-0.001	-0.024	0.006	-0.003	0.001	-0.010	-0.000
	(0.020)	(0.014)	(0.026)	(0.017)	(0.010)	(0.007)	(0.010)	(0.010)
N	3134	3263	3368	3316	3153	3309	3240	3384

	Math				Reading			
	Model 1	Model 2	Model 3	Model 4	Model 1	Model 2	Model 3	Model 4
Panel B: High-Performing Schools								
Post	-0.076***	-0.108***	-0.112***	-0.115***	-0.010	-0.036***	-0.049***	-0.029*
	(0.016)	(0.010)	(0.019)	(0.021)	(0.014)	(0.007)	(0.012)	(0.013)
Changed Principal	-0.002	-0.012	-0.003	0.010	-0.037	-0.013	0.002	0.017
	(0.018)	(0.007)	(0.015)	(0.014)	(0.019)	(0.007)	(0.011)	(0.013)
Post x Changed Principal	0.003	0.023	0.013	0.002	-0.006	0.008	-0.026	-0.026
	(0.019)	(0.013)	(0.023)	(0.025)	(0.011)	(0.007)	(0.014)	(0.015)
N	3434	3493	3344	3338	3377	3328	3436	3275

Notes: Coefficients are 2 × 2 difference-in-differences estimates from a six-year event window (-3 to 2). Changed principal is a binary indicator for schools that experienced a principal transition between year -1 and 0, where the comparison group are schools that kept the same principal for all six years. Post is a binary indicator for the post-transition period (years 0, 1, and 2). Standard errors are robust to clustering at the school-by-spell level. Model headers define the outcome (math or reading) and the residualization model (1–4). Low-performing (high-performing) schools are defined by being in the bottom (top) quartile of the three-year moving average of school performance in the pre-period.
* p < 0.05, ** p < 0.01, *** p < 0.001.

Table A.10: 2×2 Difference-in-Differences Estimates (NYC)

	Math		Reading	
	Model 1	Model 2	Model 1	Model 2
Panel A: Low-Performing Schools				
Post	0.032**	0.033***	0.036***	0.037***
	(0.010)	(0.006)	(0.009)	(0.006)
Changed Principal	-0.021*	-0.005	-0.009	-0.006
	(0.010)	(0.006)	(0.009)	(0.006)
Post x Changed Principal	0.015	-0.003	-0.004	-0.011
	(0.013)	(0.008)	(0.012)	(0.008)
N	3398	3423	3401	3396
	Math		Reading	
	Model 1	Model 2	Model 1	Model 2
Panel B: High-Performing Schools				
Post	-0.040***	-0.047***	-0.039***	-0.040***
	(0.012)	(0.007)	(0.012)	(0.006)
Changed Principal	-0.008	-0.003	-0.006	0.002
	(0.015)	(0.006)	(0.017)	(0.006)
Post x Changed Principal	-0.048***	-0.029**	-0.004	-0.009
	(0.014)	(0.009)	(0.013)	(0.007)
N	3933	3987	3927	3879

Notes: Coefficients are 2 × 2 difference-in-differences estimates from a six-year event window (-3 to 2). Changed principal is a binary indicator for schools that experienced a principal transition between year -1 and 0, where the comparison group are schools that kept the same principal for all six years. Post is a binary indicator for the post-transition period (years 0, 1, and 2). Standard errors are robust to clustering at the school-by-spell level. Model headers define the outcome (math or reading) and the residualization model (1–2). Low-performing (high-performing) schools are defined by being in the bottom (top) quartile of the three-year moving average of school performance in the pre-period.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A.11: 2×2 Difference-in-Differences Estimates (OR)

	Math				Reading			
	Model 1	Model 2	Model 3	Model 4	Model 1	Model 2	Model 3	Model 4
Panel A: Low-Performing Schools								
Post	0.059***	0.084***	0.057*	0.041	0.037**	0.062***	0.063**	0.045**
	(0.017)	(0.012)	(0.025)	(0.021)	(0.014)	(0.011)	(0.023)	(0.017)
Changed Principal	-0.011	0.010	-0.020	-0.010	-0.015	0.016	-0.027	-0.007
	(0.020)	(0.011)	(0.023)	(0.023)	(0.023)	(0.010)	(0.031)	(0.023)
Post x Changed Principal	-0.014	-0.004	0.033	0.006	-0.007	-0.013	0.011	-0.001
	(0.024)	(0.017)	(0.036)	(0.032)	(0.023)	(0.014)	(0.040)	(0.028)
N	1546	1394	541	1248	1521	1459	519	1221

	Math				Reading			
	Model 1	Model 2	Model 3	Model 4	Model 1	Model 2	Model 3	Model 4
Panel B: High-Performing Schools								
Post	-0.053***	-0.059***	-0.087***	-0.067*	-0.026	-0.041***	-0.062**	-0.092***
	(0.015)	(0.010)	(0.018)	(0.026)	(0.014)	(0.009)	(0.021)	(0.018)
Changed Principal	0.012	-0.005	0.063*	-0.011	0.028	-0.001	0.027	-0.003
	(0.035)	(0.014)	(0.030)	(0.027)	(0.030)	(0.011)	(0.024)	(0.024)
Post x Changed Principal	0.010	0.013	-0.039	-0.028	-0.022	-0.009	-0.009	0.009
	(0.032)	(0.015)	(0.039)	(0.034)	(0.028)	(0.014)	(0.031)	(0.026)
N	1503	1442	545	1395	1545	1465	531	1291

Notes: Coefficients are 2 × 2 difference-in-differences estimates from a four-year event window (-2 to 1). We limit the event window to four years in OR due to limited sample size with six year windows. Changed principal is a binary indicator for schools that experienced a principal transition between year -1 and 0, where the comparison group are schools that kept the same principal for all six years. Post is a binary indicator for the post-transition period (years 0 and 1). Standard errors are robust to clustering at the school-by-spell level. Model headers define the outcome (math or reading) and the residualization model (1–4). Low-performing (high-performing) schools are defined by being in the bottom (top) quartile of the two-year moving average of school performance in the pre-period.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A.12: Variance Decomposition Results with AR(1) Autocorrelated Errors (Math)

	Model 0	Model 1	Model 2	Model 3	Model 4
Panel A: Tennessee					
Random Effects Parameters (SD)					
School	0.412	0.297	0.129	0.171	0.167
Principal	0.000	0.000	0.048	0.054	0.042
Residual	0.215	0.209	0.170	0.185	0.179
AR(1) Parameters					
Correlation $(t-1)$	0.590	0.594	0.345	0.414	0.531
Panel B: New York City					
Random Effects Parameters (SD)					
School	0.500	0.225	0.111		
Principal	0.000	0.000	0.046		
Residual	0.220	0.183	0.118		
AR(1) Parameters					
Correlation $(t-1)$	0.789	0.710	0.245		
Panel C: Oregon					
Random Effects Parameters (SD)					
School	0.362	0.310	0.150	0.212	0.197
Principal	0.000	0.000	0.052	0.042	0.000
Residual	0.182	0.175	0.146	0.169	0.158
AR(1) Parameters					
Correlation $(t-1)$	0.591	0.588	0.215	0.414	0.530

Notes: Cells report standard deviations of variance components and percentage of overall variance explained from Equation 9. Model 0 uses raw test scores, Model 1 adjusts for student demographic characteristics, Model 2 adds cubic polynomials of lagged-test scores and attendance, Model 3 restricts to students in first year in school, Model 4 uses prior-school outcomes. Models 3 and 4 not estimated for NYC because we do not observe year of first enrollment. All models include grade and year fixed effects. Demographic covariates include prior grade retention, gender, race/ethnicity, disability status, 504 plan designation, participation in migrant or Indian education program and the school averages of the preceding characteristics. All samples restricted to observations with at least 25 students in each school-by-year cell.

Table A.13: Variance Decomposition Results with Autocorrelated Errors (Reading)

	Model 0	Model 1	Model 2	Model 3	Model 4
Panel A: Tennessee					
Random Effects Parameters (SD)					
School	0.429	0.288	0.114	0.162	0.139
Principal	0.000	0.000	0.018	0.000	0.000
Residual	0.148	0.139	0.106	0.125	0.115
AR(2) Parameters					
Correlation $(t-1)$	0.528	0.522	0.179	0.266	0.431
Correlation $(t-2)$	0.076	0.075	0.136	0.108	0.120
Panel B: New York City					
Random Effects Parameters (SD)					
School	0.486	0.225	0.110		
Principal	0.000	0.000	0.005		
Residual	0.194	0.156	0.108		
AR(2) Parameters					
Correlation $(t-1)$	0.670	0.582	0.173		
Correlation $(t-2)$	0.115	0.116	0.172		
Panel C: Oregon					
Random Effects Parameters (SD)					
School	0.339	0.273	0.132	0.168	0.162
Principal	0.000	0.000	0.000	0.020	0.000
Residual	0.159	0.151	0.137	0.148	0.132
AR(2) Parameters					
Correlation $(t-1)$	0.525	0.505	0.233	0.306	0.403
Correlation $(t-2)$	0.076	0.091	0.157	0.120	0.074

Notes: Cells report standard deviations of variance components and percentage of overall variance explained from Equation 9. Model 0 uses raw test scores, Model 1 adjusts for student demographic characteristics, Model 2 adds cubic polynomials of lagged-test scores and attendance, Model 3 restricts to students in first year in school, Model 4 uses prior-school outcomes. Models 3 and 4 not estimated for NYC because we do not observe year of first enrollment. All models include grade and year fixed effects. Demographic covariates include prior grade retention, gender, race/ethnicity, disability status, 504 plan designation, participation in migrant or Indian education program and the school averages of the preceding characteristics. All samples restricted to observations with at least 25 students in each school-by-year cell.

Table A.14: First-Differences Estimates Predicting Mean Test Score Residuals (Model 2, TN)

	Δ Math Residuals				Δ Reading Residuals			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Δ Math Value-Added	0.280***	0.284***		0.214***		0.030		0.025
	(0.024)	(0.027)		(0.025)		(0.017)		(0.017)
Δ Reading Value-Added		0.092		0.084	0.164***	0.172***		0.123***
		(0.053)		(0.052)	(0.035)	(0.039)		(0.036)
Δ Teacher Experience			0.001	-0.001			-0.000	-0.001
			(0.001)	(0.002)			(0.001)	(0.001)
N	10415	8607	14205	8594	10316	8607	14246	8594

Notes: Coefficients shown are from first differences models where the dependent variable is defined by the header. Columns 1, 2, and 4 (5, 7, and 8) are weighted by the number of teachers in the school-by-year cell with non-missing math (reading) VA. Teacher-level VA estimates are produced using the drift-adjusted framework outlined in Chetty, Friedman, and Rockoff (2014a), where we predict VA in year t only using test score residuals from when a teacher worked in a different school. Columns 3 and 7 are weighted by the total number of teachers in the school-by-year cell. Heteroskedasticity-robust standard errors shown in parentheses.

* p < 0.05, ** p < 0.01, *** p < 0.001.

Table A.15: First-Differences Estimates Predicting Mean Test Score Residuals (Model 2, NYC)

	Δ Math Residuals				Δ Reading Residuals			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Δ Math Value-Added	0.083***	0.096***		0.067***		0.043*		0.035*
	(0.019)	(0.021)		(0.018)		(0.018)		(0.017)
Δ Reading Value-Added		-0.022		-0.018	-0.009	-0.027		-0.036
		(0.024)		(0.022)	(0.021)	(0.023)		(0.021)
Δ Teacher Experience			-0.001	-0.001			0.003**	0.003**
			(0.001)	(0.001)			(0.001)	(0.001)
N	15426	14959	16588	15083	15326	14965	16590	15083

Notes: Coefficients shown are from first differences models where the dependent variable is defined by the header. Columns 1, 2, and 4 (5, 7, and 8) are weighted by the number of teachers in the school-by-year cell with non-missing math (reading) VA. Teacher-level VA estimates are produced using the drift-adjusted framework outlined in Chetty, Friedman, and Rockoff (2014a), where we predict VA in year t only using test score residuals from when a teacher worked in a different school. Columns 3 and 7 are weighted by the total number of teachers in the school-by-year cell. Heteroskedasticity-robust standard errors shown in parentheses.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A.16: First-Differences Estimates Predicting Mean Test Score Residuals (Model 2, OR)

	Δ Math Residuals				Δ Reading Residuals			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Δ Math Value-Added	0.209***	0.194**		0.153**		0.015		0.029
	(0.047)	(0.062)		(0.053)		(0.061)		(0.052)
Δ Reading Value-Added		0.184**		0.175**	0.204***	0.213**		0.165**
		(0.064)		(0.058)	(0.058)	(0.065)		(0.057)
Δ Teacher Experience			0.000	0.005*			-0.001	0.005*
			(0.001)	(0.002)			(0.001)	(0.002)
N	2057	1426	9866	1427	1747	1423	9864	1427

Notes: Coefficients shown are from first differences models where the dependent variable is defined by the header. Columns 1, 2, and 4 (5, 7, and 8) are weighted by the number of teachers in the school-by-year cell with non-missing math (reading) VA. Teacher-level VA estimates are produced using the drift-adjusted framework outlined in Chetty, Friedman, and Rockoff (2014a), where we predict VA in year t only using test score residuals from when a teacher worked in a different school. Columns 3 and 7 are weighted by the total number of teachers in the school-by-year cell. Heteroskedasticity-robust standard errors shown in parentheses.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

B Data Description

Table B.1: Descriptive Statistics (Tennessee)

	Math Sample		Reading Sample		Attend Sample	
	Mean	SD	Mean	SD	Mean	SD
Students						
Asian	0.02		0.02		0.02	
American Indian	0.00		0.00		0.00	
Black	0.24		0.24		0.24	
Hispanic	0.07		0.07		0.08	
Pacific Islander	0.00		0.00		0.00	
White	0.66		0.67		0.66	
Qualifies for FRPL	0.49		0.48		0.49	
Enrolled in Special Education	0.10		0.11		0.13	
English Learner Classification	0.03		0.02		0.03	
Standardized Math Score	-0.01	0.99	0.01	1.00		
Standardized Reading Score	-0.01	0.99	0.01	0.99		
Proportion Days Absent	0.05	0.05	0.05	0.06	0.05	0.06
Standardized Math Score (prior-year)	0.01	0.96	0.04	0.98		
Standardized Reading Score (prior-year)	-0.00	0.97	0.02	0.97		
Proportion Days Absent (prior-year)	0.05	0.05	0.05	0.05	0.05	0.05
Missing Prior-year Math Score	0.07		0.06			
Missing Prior-year Reading Score	0.06		0.06			
Missing Prior-year Absence Rate	0.04		0.04		0.03	
Sample Size (Student-by-Year)	5,225,333		5,841,584		9,991,519	
Unique Students	1,427,053		1,488,638		1,945,046	
Principals						
Female	0.55		0.55		0.56	
Black	0.19		0.19		0.19	
White	0.81		0.81		0.81	
Other Race/Ethnicity	0.00		0.00		0.00	
Age	49.75	8.95	49.74	8.95	49.68	8.99
Years of Experience (total)	22.33	9.24	22.32	9.25	22.22	9.29
Years of Experience (principal)	4.87	3.80	4.86	3.80	4.90	3.86
Years in Current School (principal)	3.74	3.43	3.74	3.42	3.77	3.48
Elementary School	0.57		0.57		0.59	
Middle School	0.20		0.20		0.19	
High School	0.19		0.19		0.18	
Other Level School	0.04		0.04		0.04	
Sample Size (Principal-by-Year)	17,553		17,577		19,867	
Unique Principals	3,925		3,925		4,095	

Table B.2: Descriptive Statistics (New York City)

	Math Sample		Reading Sample	
	Mean	SD	Mean	SD
Students				
Female	0.49		0.49	
Asian	0.14		0.14	
Black	0.31		0.31	
Hispanic/Latino	0.40		0.39	
White	0.15		0.15	
Other Race/Ethnicity	0.01		0.01	
English Learner Classification	0.12		0.09	
Qualifies for FRPL	0.82		0.82	
Enrolled in Special Education	0.16		0.17	
Standardized Math Score	-0.00	1.00	0.03	0.99
Standardized Reading Score	0.00	1.00	0.00	1.00
Standardized Math Score (prior-year)	0.04	0.97	0.06	0.97
Standardized Reading Score (prior-year)	0.04	0.97	0.04	0.97
Missing Prior-year Math Score	0.08		0.06	
Missing Prior-year Reading Score	0.12		0.09	
Sample Size (Student-by-Year)	6,194,478		5,976,223	
Unique Students	1,834,499		1,773,424	
Principals				
Female	0.71		0.71	
Black	0.27		0.27	
White	0.48		0.48	
Hispanic/Latino	0.07		0.07	
Other Race/Ethnicity	0.03		0.03	
Missing Race/Ethnicity	0.15		0.15	
Age	50.57	8.36	50.57	8.36
Years of Experience (total)	26.42		26.42	
Years of Experience (principal)	4.84	4.41	4.84	4.41
Years in Current School (principal)	4.60	4.41	4.60	4.41
Sample Size (Principal-by-Year)	18,238		18,240	
Unique Principals	3,200		3,201	

Table B.3: Descriptive Statistics (Oregon)

	Math Sample		Reading Sample		Attend Sample	
	Mean	SD	Mean	SD	Mean	SD
Students						
Asian/Pacific Islander	0.07		0.07		0.07	
American Indian/Alaska Native	0.02		0.02		0.02	
Black	0.02		0.02		0.03	
Hispanic/Latino	0.21		0.22		0.21	
White	0.65		0.65		0.65	
Other Race/Ethnicity	0.03		0.03		0.03	
Qualifies for FRPL	0.51		0.51		0.50	
Enrolled in Special Education	0.13		0.13		0.14	
Limited English Proficiency	0.07		0.07		0.08	
504 Plan Designation	0.02		0.02		0.02	
Migrant Designation	0.02		0.02		0.02	
Indian Education Designation	0.01		0.01		0.01	
Standardized Math Score	0.04	0.99				
Standardized Reading Score			0.04	0.99		
Proportion Days Absent					0.06	0.07
Standardized Math Score (prior-year)	0.04	0.99	0.04	0.99		
Standardized Reading Score (prior-year)	0.03	0.99	0.03	0.99		
Proportion Days Absent (prior-year)	0.05	0.06	0.05	0.06	0.06	0.06
Missing Prior-year Math Score	0.10		0.13			
Missing Prior-year Reading Score	0.14		0.10			
Missing Prior-year Absence Rate	0.03		0.02		0.03	
Sample Size (Student-by-Year)	2,874,460		2,830,334		5,419,600	
Unique Students	846,570		839,055		1,134,496	
Principals						
Female	0.50		0.50		0.50	
American Indian	0.01		0.01		0.01	
Asian/Pacific Islander	0.02		0.02		0.02	
Black	0.02		0.02		0.02	
Hispanic/Latino	0.05		0.05		0.05	
Multi-Racial	0.02		0.02		0.02	
White	0.86		0.86		0.86	
Other Race/Ethnicity	0.03		0.03		0.03	
Age	47.95	8.10	47.95	8.10	48.00	8.15
Years of Experience (total)	19.33	8.48	19.33	8.48	19.29	8.53
Years of Experience (principal)	2.82	2.63	2.81	2.63	2.80	2.63
Years in Current School (principal)	1.99	2.10	1.99	2.10	1.97	2.10
Elementary School	0.50		0.50		0.50	
Middle School	0.17		0.17		0.17	
High School	0.16		0.16		0.16	
Other Level School	0.16		0.16		0.16	
Sample Size (Principal-by-Year)	11,815		11,819		12,449	
Unique Principals	2,651		2,653		2,757	

B.1 Tennessee

Data Construction and Sample Restrictions. The Tennessee data are constructed from yearly base datasets of staff and students, respectively. The staff data are available beginning in the 2001–02 school year and include demographic, position/assignment, and salary information for all individuals working in a K–12 public school in Tennessee. These yearly data files allow us to identify the principal and teachers working in a particular school. In a small percentage of schools each year (roughly 5%), there is insufficient information to reliably identify a single principal either because no staff member identified as a principal is working in the school, or because there are many identified principals in one school. We drop these cases from our analytic dataset. Between 2006–07 and 2018–19, we observe 22,174 school-by-year observations with an identified principal. This baseline sample is further reduced by the availability of student test score and attendance data, described below.

The student-level data are first available in 2006–07 and include information about each student's demographics, specific dates of enrollment and withdrawal at each school they attended during the year, daily attendance records, and scores on end-of-year exams. Test score records include a school identifier, which is how we link students to schools in a particular year. Each student is only linked to one school in a given year based on this information. Test scores include statewide end-of-year exams in math and reading for grades 3–8 and end-of-course exams for high school students and advanced middle school students. End-of-course exams in math subjects include algebra I, (2007–2019), algebra II (2012–2019), geometry I (2016–2019), and integrated math I, II, and III (separate exams for each, 2016–2019). End-of-course exams in reading subjects include English I (2007–2019), English II, (2007–2019), and English III (2012–2018). Prior to 2011–12, 7th and 8th grade students who were enrolled in Algebra I courses took the end-of-course exams.

For the math test score sample, we begin with 6,807,025 million observed student-year-exam observations. We then make the following successive restrictions: (1) we drop all observations from 2006–07 (495,346) because there are no prior-year test scores. (2) We then eliminate 3rd grade students (811,341) because they have no prior test score. (3) We then drop 15,596 observations where there were inconsistencies in the student's assigned grade based on enrollment data and test score data. (4) We then drop 131,713 observations for middle or elementary school students who took EOC exams in Algebra I. These students also took their respective end-of-grade exams and thus are retained in the sample, but we avoid duplication. This leaves us with 5,353,029 student-by-year observations. (5) We then drop 4th grade students in the 2016–17 school year (77,035) because they have no prior-year test score. This arises because Tennessee cancelled statewide testing in grades 3–8 for the 2015–16 school year. For students in grades 5–8, we use their twice-lagged test score as the prior-year test score for 2016–17. (6) We then drop 46,556 observations with missing demographic information. (7) Finally, we drop 4,105 observations in school-by-year cells with fewer than 25 observations. This leaves us with a final analytic sample of 5,225,333 student-by-year observations for math.

For reading, we follow the same steps beginning from an initial sample of 7,359,843 million observed student-year-exam observations. The reductions at each step are: (1) 557,551; (2) 809,475; (3) 15,567; (4) 1,591 (students in grade 8 or below taking an EOC English exam); (5) 76,759; (6) 53,325; and (7) 3,991. This leaves us with a final analytic sample of 5,841,584 student-by-year observations for reading. The larger sample size for reading is due to more student taking an end-of-course exam in reading than math.

For attendance, we begin with 14,048,979 student-by-school-by-year observations. We first restrict to students who attended a single school for at least 110 instructional days, which drops 1,884,674 observations. We then drop students who are recorded as being enrolled in two or more schools for at least 110 instructional days (276,541). We then drop 954,910 kindergarten students and 821,679 students in 2006–07 (no prior-year attendance). Finally, we drop 114,891 students with missing demographics and 4,765 students in school-by-year cells with fewer than 25 observations. This leaves us with a final analytic sample of 9,991,519 students for attendance.

Measures. The Tennessee data include information on the following student demographic characteristics: gender, race/ethnicity, parental income (as measured by eligibility for free- or reduced-price lunch), special education status, and English learner status. We include these and school averages of the same variables as covariates in our models that residualize student test scores and attendance rates.

B.2 New York City

Data Construction and Sample Restrictions. The New York City data used in these analyses emerge from yearly administrative datasets that contain, in separate files, principal, teacher, and student records from the 1999 academic year to the 2017 academic year. Staff (principals and teachers) and students are linked across these files by de-identified staff and school identifiers and academic years – variables that appear across the respective datasets. We drop student and staff records that are missing any of the relevant identifiers.

Students in our analytic sample must be in 3rd-8th grade and have Math and/or reading test score outcomes. We require that students have complete demographic data, including information on gender, race, English Learner status, Free and Reduced Price Lunch status, and disability status. We eliminate those students who have math and reading test results from different schools (13,094 observations) as well as duplicate records (1,692,116 observations). To be included in the analysis, student-by-year records should also contain current and prior-year outcomes in the respective tested subject. As a result, we eliminate 780,129 student-by-year observations from the 1999 school year (the first year of data we have), 1,304,654 3rd grade observations in math (the first grade with test outcomes), and 1,249,658 3rd grade observation in reading. Finally, we drop 3,596 (math) and 4,328 (reading) students in school-by-year cells with fewer than 25 observations. We are left with 6,194,478 student observations in our math analytic sample and 5,976,223 student observations in our reading sample.

Our sample includes only those schools that have one principal of record. Principals in our analytic sample must also have continuous tenure. We eliminate those principal-by-school spells that do not – i.e. those instances where a year (or more) of data is missing for a school, but the same principal shows up before and after the break in data – losing 408 observations. Upon connecting the principal data with student data for students in tested grades and subjects, and after imposing the aforementioned analytic data restrictions, we end up with 18,238 principal observations in our math analytic file and 18,240 principal observations for our reading sample.

The teacher analytic sample is restricted to teachers who are labeled "paid, regular teachers." We also only include teachers who are rostered to students for purposes of calculating value-added (i.e. students in tested grades and subjects after applying the aforementioned restrictions).

Measures. The NYC data include information on the following student demographic characteristics: gender, race/ethnicity, parental income (as measured by eligibility for free- or reduced-price lunch), special education status, and English learner status. We include these and school averages of the same variables as covariates in our models that residualize student test scores.

B.3 Oregon

Data Construction and Sample Restrictions. We construct the Oregon data from three separate data sources that describe (a) students' demographic and school enrollment status; (b) students' test scores; and (c) all staff employed in the Oregon public school system. We link principals and students through students' attended school of record and principals' assigned institutional organization. To appear in our sample, students must have attended a school for at least 110 days in a given year and have a current-year outcome. We assign students with missing prior-year tests a prior-year score of 0 and include an indicator for missing prior score. A very small number of our observations have missing demographic information with almost all of the missingness in the years prior to 2009–10 (between 0.05 percent and 2 percent of our observations have missing demographic information, depending on the variable). We assign these observations values of 0 for that demographic variable and create indicators for missing demographic information which we use in the residualization process. All results are robust to excluding observations with missing demographics.

We restrict our test-score samples to grades 4–12 and our attendance sample to grades 1–12, so that we can observe prior outcomes. We require principals to either be principal of only a single school in a year or to have the highest FTE of any educator assigned as a principal to that school in that year. This represents 96 percent of principal-year observations. Generally, student mobility across schools and from outside the public school system has a substantial effect on our sample, whereas the other restrictions are marginal.

After eliminating students' secondary schools of attendance in a given year, students recorded as having zero days present, and a very small number of students recorded as under 4 or over 21 (813 student-year observations), we have samples of 3,140,724; 3,110,617 and 6,438,821 student-year observations in our math, reading and attendance sample, respectively. We make the following additional restrictions in sequence in math: drop 90,111 observations from 2006–07 as we do not observe prior test scores, drop 170,841 student-year observations with less than 110 days attendance (present or absent recorded) in a single school; drop 5,312 student-year observations in school-by-year cells with fewer than 25 students. This results in our *final analytic math sample of 2,874,460 student-year observations*. We make the following additional restrictions in sequence in reading: drop 101,221 observations from 2006–07 as we do not observe prior test scores, drop 173,781 student-year observations with less than 110 days attendance (present or absent recorded) in a single school; drop 5,281 student-year observations in school-by-year cells with fewer than 25 students. This results in our *final analytic reading sample of 2,830,334 student-year observations*. We make the following additional restrictions in sequence in attendance: drop 482,381 observations from 2006–07 as we do not observe prior test scores, drop 482,381 student-year observations with less than 110 days attendance (present or absent recorded) in a single school; drop 1,454 student-year observations in school-by-year cells with fewer than 25 students. This results in our *final analytic attendance sample of 5,419,600 student-year observations*.

These students are, in turn, linked with over 2,650 and 2,750 unique principals in our test-score and attendance samples in Oregon, respectively.

Measures. The Oregon data include information on the following student demographic characteristics: gender, race/ethnicity, parental income (as measured by eligibility for free- or reduced-price lunch), special education status, limited English proficiency, 504 plan designation, and participation in migrant or Indian education programming. We also include indicators for missing demographic variables. We include these and school averages of the same variables as covariates in our models that residualize student test scores and attendance rates.

Statewide teacher-student linkages are only possible in Oregon starting in the 2013–14 school year. Thus, our mechanism results that rely on teacher-value-added estimates draw on only the final six years of our sample and only on teachers who teach math or reading.