Rethinking Principal Effects on Student Outcomes

Brendan Bartanen
University of Virginia

Aliza N. Husain
Pivot Learning

David D. Liebowitz
University of Oregon

November 2023

Abstract

School principals are viewed as critical actors 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. As commonly implemented, value-added models misattribute to principals changes in student performance caused by unobserved time-varying factors over which principals exert minimal control, leading to biased estimates of individual principals' effectiveness and an overstatement of the magnitude of principal effects. Based on our framework, which better accounts for bias from time-varying factors, we find that little of the variation in student test scores or attendance is explained by persistent effectiveness differences between principals. Across contexts, the estimated standard deviation of principal value-added is roughly 0.03 student-level standard deviations in math achievement and 0.01 standard deviations in reading.

***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 (e.g., Branch, Hanushek, and Rivkin 2012; Coelli and Green 2012; Dhuey and Smith 2018). 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.[1] 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 high-performing principals using the outcomes of students and schools? How important is

1. 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 better 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.

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 typical 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.

Consistent with prior work, we first demonstrate that between-principal differences in school performance (as measured by student outcomes) are substantial. In baseline variance

decomposition models that attribute to principals all persistent or semi-persistent changes in school performance, a one standard deviation increase in the principal effectiveness distribution translates to an increase of roughly 0.08 student-level standard deviations (*SD*) in math achievement and 0.05 *SD* in reading. Consistently across all three datasets, however, we show that cross-year correlations in school performance within and across principals are very similar, such that only a small fraction—19% in math and 5% in reading—of the between-principal variance is truly a function of principal effectiveness. Once accounting for time-varying factors that principals do not control, we estimate the magnitude of principal effects to be approximately 0.03 *SD* in math and 0.01 *SD* in reading.

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, which leads to an upward bias in the estimated magnitude of principal effects. 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 relatively little 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 primary contribution is to suggest that the true magnitude of principals' effects on student test scores is substantially smaller than previously believed based on the existing literature. A survey of U.S.-based studies that estimate principal VA (all relying on identifi-

cation from within-school differences across principals) finds an unweighted average of 0.13 *SD* in math and 0.09 *SD* in reading (Grissom, Egalite, and Lindsay 2021). However, there is substantial variability across studies. Using data from Pennsylvania, Chiang, Lipscomb, and Gill (2016) estimate principal effects of 0.14 *SD* (math) and 0.11 *SD* (reading). Dhuey and Smith (2018) find corresponding estimates of 0.17 *SD* and 0.12 *SD* in North Carolina. In Tennessee, Bartanen (2020) finds 0.20 *SD* and 0.10 *SD*. Examining math achievement only in Texas and Chicago, respectively, Branch, Hanushek, and Rivkin (2012) and Laing et al. (2016) estimate principal effects between 0.05–0.11 *SD* and 0.04–0.08 *SD*.

Existing studies likely overstate the true magnitude of principal effects for two reasons. First, Bartanen and Husain (2022) demonstrate how limited mobility of principals among schools leads to variance inflation in models with principal and school fixed effects. The current paper goes a step further in interrogating the reliability and validity of the identifying variation in principal VA models. Specifically, we demonstrate how these models erroneously attribute to principals the effects of time-varying, unobserved school heterogeneity, which creates upward bias in the estimated variance of principal effects. Our estimates of 0.03 *SD* in math and 0.01 *SD* in reading overcome both of these sources of upward bias and, accordingly, are considerably smaller than prior work.

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 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 school personnel, 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; Cullen et al. 2021). 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.

Researchers generally specify VA models in an *ad-hoc* fashion and, despite not having a structural interpretation, view the resulting estimates as potentially informative of educators' causal effects (Koedel, Mihaly, and Rockoff 2015; Rubin, Stuart, and Zanutto 2004). 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.[2] 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.[3] 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.[4] 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 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

2. 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.

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

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

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 2022).

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. If so, existing approaches to estimate principal effects may not entirely resolve the challenge of separating the principal's contribution from uncontrollable school factors.

To introduce additional flexibility into the conceptual model relating principal effectiveness to changes in school performance, 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. That is, we allow both the effectiveness of principals and other school factors to fluctuate over time. Note that $\text{cov}(\delta^F, \delta^D) = \text{cov}(\mu^F, \mu^D) = 0$, by construction, such that these dynamic components, δ^D and μ^D, are defined as deviations from mean school- and principal-performance across years (δ^F and μ^F, respectively).

While Equation 2 adds flexibility, it requires certain assumptions to remain empirically tractable. Specifically, we follow Chetty, Friedman, and Rockoff (2014a) in specifying that

both δ^D and μ^D fluctuate stochastically over time according to a stationary process:

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

where x denotes the difference between years for a given t. Here, the stationarity assumption means 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). This reduces the number of parameters to be estimated and also implies that the dynamic components (μ^D and δ^D) are orthogonal. As we describe further below, the stationarity assumption allows us to obtain an estimate of the variance of principals' contributions to school performance by comparing the cross-year stability of school performance within and across principals. We also perform various checks to probe the plausibility of this assumption.

In practical terms, μ^D in Equation 2 makes the model relating principal effectiveness to school performance more plausible by allowing 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 provides variation we can leverage to 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 of the magnitude of principals' contributions to school performance.[5] To see this, we can write these correlations as:

$$r_x^{\text{SamePrin}} = \frac{\sigma^2_{\mu^F} + \sigma_{\mu^D_x} + \sigma^2_{\delta^F} + \sigma_{\delta^D_x}}{\sigma^2_Y}, \tag{4}$$

$$r_x^{\text{DiffPrin}} = \frac{\sigma^2_{\mu^F} + \sigma_{\mu^D_x} + \sigma_{\delta^F_x}}{\sigma^2_Y},$$

$$\text{where } \sigma_{\delta^F_x} = \text{cov}(\delta^F_{j(s,t)}, \delta^F_{k(s,t+x)}) < \sigma^2_{\delta^F} \text{ for } j \neq k$$

We assume that $\text{cov}(\mu^D_x, \delta^D_x) = \text{cov}(\mu^D_x, \delta^F_x) = \text{cov}(\mu^F_x, \delta^D_x) = 0$, meaning that the dynamic components of δ and μ are uncorrelated with each other and with the respective fixed components.[6] We also assume that $\text{cov}(\delta^D_{jt(s,t)}, \delta^D_{kt(s,t+x)}) = 0$, meaning that it is not the case that

5. This framework is similar to that used by Branch, Hanushek, and Rivkin (2012)—and extended on by Laing et al. (2016)—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. This provides an estimate of the within-school variance in principal quality. 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. We show a direct comparison of their approach and ours in Appendix Table A.4.

6. Conceptually, $\text{cov}(\mu^D_x, \delta^D_x) \neq 0$ would indicate systematic sorting of principals to schools on the basis of temporary fluctuations in principal and school effectiveness, which seems unlikely, particularly given that the majority of new-to-school principals have no prior principal experience. $\text{cov}(\mu^D_x, \delta^F_x) \neq 0$ or $\text{cov}(\mu^F_x, \delta^D_x) \neq 0$ are perhaps more plausible, though still unlikely. $\text{cov}(\mu^D_x, \delta^F_x) \neq 0$ could arise from compensatory assignment practices if, for instance, districts aim to re-assign effective principals to schools that are struggling (relative to their typical performance). However, that would require reliable information about principals' true effectiveness and a concerted reassignment strategy, which seems unlikely given high rates of principal

certain schools are repeatedly led by principals whose performance is temporarily higher or lower than their typical performance. However, we allow for the possibility of nonrandom sorting of principals to schools on the basis of their fixed components. This is represented by $\sigma_{\delta_x^F}$ 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. We discuss the implications of this sorting below.

The difference between r_x^{SamePrin} and r_x^{DiffPrin} provides insight about the extent to which school performance is driven by principals versus school-level factors that principals cannot control. Following from Equation 4, we can write:

$$r_x^{\text{SamePrin}} - r_x^{\text{DiffPrin}} = \frac{\sigma^2_{\delta^F} + \sigma_{\delta_x^D} - \sigma_{\delta_x^F}}{\sigma^2_Y} \tag{5}$$

We can then use $\sigma^2_Y(r_x^{\text{SamePrin}} - r_x^{\text{DiffPrin}})$ to calculate $\sigma^2_{\delta^F} + \sigma_{\delta_x^D} - \sigma_{\delta_x^F}$, which is the lower-bound estimate of the variance of principal effects, where the true variance is $\sigma^2_{\delta^F} + \sigma_{\delta_x^D}$. This comparison is similar to a difference-in-differences logic. That is, r_x^{DiffPrin} acts as a counterfactual: how stable would school performance have been if the school did not keep the same principal? This allows us to understand how much of the stability indicated by r_x^{SamePrin} is driven by principal effectiveness as opposed to factors that principals cannot control. If higher-quality principals cause improved school performance, then $\sigma^2_{\delta^F} + \sigma_{\delta_x^D} - \sigma_{\delta_x^F} > 0$ 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_p^F relative to δ_{pt}^D). 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.

We characterize $r_x^{\text{SamePrin}} - r_x^{\text{DiffPrin}}$ as a lower-bound estimate because we anticipate that

attrition. $\text{cov}(\mu_x^F, \delta_x^D) \neq 0$ would imply that principals sort to high- or low-performing schools on the basis of fluctuations in their average effectiveness. Again, this could arise if districts are systematically reassigning temporarily high- or low-performing principals to certain types of schools.

there is some degree of non-random (positive) sorting, such that $\sigma_{\delta_x^F} > 0$ and $r_x^{\text{SamePrin}} - r_x^{\text{DiffPrin}} < \sigma_{\delta^F}^2 + \sigma_{\delta_x^D}$. In other words, this difference in correlations will understate the magnitude of principal effects if certain schools are systematically more likely to be led by higher-quality principals.[7] Based on Equation 4 and assuming (conservatively) sorting patterns are static, we can write this bias as:

$$\sigma_{\delta_x^F} = \rho_\delta \sigma_{\delta^F}^2 \tag{6}$$

to see that $\sigma_{\delta_x^F}$ is increasing in the school-level intraclass correlation of principal quality (ρ_δ) and the variance of the stable component principal effects ($\sigma_{\delta^F}^2$).[8] Equation 6 helps to establish that $r_x^{\text{SamePrin}} - r_x^{\text{DiffPrin}}$ is likely a useful lower-bound estimate of magnitude of principal effects because any bias scales with $\sigma_{\delta^F}^2$, which we further demonstrate in Section 5.4.1.[9]

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,

7. Note that the converse also holds; $r_x^{\text{SamePrin}} - r_x^{\text{DiffPrin}}$ becomes an upper-bound estimate if $\sigma_{\delta_x^F} < 0$, meaning that schools are more likely to replace a high-quality leader with a low-quality leader (and vice-versa). Such a scenario could arise if school districts assign principals in a compensatory equalizing fashion. We do not expect that this is a common practice.

8. Specifically, this assumption means that $\sigma_{\delta_x^F} = \sigma_{\delta^F}$, meaning that the correlation of δ^F across principals does not vary by x, which is the number of years between when they enter the school. This is likely a conservative assumption in terms of the magnitude of sorting bias. If $\sigma_{\delta_x^F}$ deteriorates with x, we will overstate the magnitude of sorting bias.

9. In fact, we show in Appendix Figure C.1 that except under fairly extreme sorting scenarios, a value of $r_x^{\text{SamePrin}} - r_x^{\text{DiffPrin}}$ that is small in magnitude cannot correspond to a large true magnitude of δ. As one example, if $\rho_\delta = 0.5$ (i.e., 50% of the variance in δ^F is within schools), the true magnitude of principal effects is 0.1 *SD* (i.e., a 1 *SD* increase in principal quality raises student test scores by 0.1 SD), and 75% of the principal effect is fixed, the estimated *SD* using $r_x^{\text{SamePrin}} - r_x^{\text{DiffPrin}}$ is 0.079, or a downward bias of 0.021 SD.

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. Thus, 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.[10] We also examine their daily attendance rates in auxiliary models. Within each dataset, these student outcomes 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.[11] 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.

10. 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 11^{th}-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.

11. 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.

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:

$$\begin{aligned} Y_{ist} &= \beta \mathbf{X}_{ist} + \gamma_{s,p} + \epsilon_{ist} \\ Y^*_{ist} &= Y_{ist} - \hat{\beta} \mathbf{X}_{ist} \end{aligned} \tag{7}$$

We estimate β using within principal-by-school spell variation by including a principal-by-school fixed effect ($\gamma_{s,p}$), which avoids overstating the impact of observables on student test scores due to a potential correlation between $\mathbf{X}_{ist}$ and school or principal 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 cells where fewer than 25 students contribute to $\bar{Y}^*_{st}$.[12] 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:

$$\bar{Y}^*_{spt} = \theta_s + \theta_{s,p} + \epsilon_{spt} \tag{8}$$

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 variance of $\theta_{s,p}$, which is the magnitude of the variation attributed to principals.[13] Equation 8 follows the

12. 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.

13. 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 8 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 2022). Additionally, there are relatively few principals

logic of prior principal VA studies that attribute persistent or semi-persistent differences in within-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 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.[14] 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.

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 in Appendix G.

14. In Tennessee and New York City, these student characteristics include gender, race/ethnicity, family 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.

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.[15]

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 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

15. 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.

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 7, 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 is greater than 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.

We can understand the potential strengths and weaknesses of different approaches for residualizing student test scores by contrasting the results across model specifications. Com-

paring 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, which suggests a large role of estimation or measurement error in teacher VA, but may also reflect instability in true performance. 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 differs 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.

The autocorrelation vectors in Figure 1 demonstrate that school performance has a substantial dynamic component. This is perhaps conceptually intuitive, 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 $t+x$ on subsamples of years where the principal in year t is either the same or different as year $t+x$.

We find only small differences in school performance autocorrelations comparing within-principal spells, as compared to across principals. For each value of x up to 10 years, we plot these correlations in Figure 2 and provide more detailed information in Tables 2 (math) and 3 (reading), including sample sizes, $r^{\text{SamePrin}} - r^{\text{DiffPrin}}$, p-values for the null hypothesis of no difference, and the implied standard deviation of principal effects based on $\sigma^2_Y(r^{\text{SamePrin}} - r^{\text{DiffPrin}})$. 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.[16] While the correlations within the same principal tend to be slightly larger in magnitude, the pattern of decreasing correlations in school performance across time is largely *not* explained by principal transitions.

Based only on the year immediately after a principal transition ($r_1^{\text{SamePrin}} - r_1^{\text{DiffPrin}}$), the estimated *SD* of principal VA in math is 0.035, 0.030, and 0.041 in Tennessee, NYC, and Oregon, respectively. While statistically significant, these estimates are substantially smaller than both those reported in nearly all prior studies and based on the variance decomposition in Table 1.[17] The bottom row of each panel also shows pooled *SD* estimates based on a weighted average of $r_x^{\text{SamePrin}} - r_x^{\text{DiffPrin}}$ across x, which are similar to the $x = 1$ estimates. For reading, $r_x^{\text{SamePrin}} - r_x^{\text{DiffPrin}}$ is close to zero for any x across all three contexts; our pooled estimates of the *SD* principal effects are 0.016, 0.000, and 0.018 *SD*, respectively.[18]

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

17. 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.4. 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. In particular, 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.

18. Our results also diverge sharply from Bartanen (2020), who applies a similar model using Tennessee data and finds large differences across principals for both test scores and attendance. While Bartanen

In general, our *SD* estimates are relatively stable for $x < 5$, which is where we also have greater precision. At higher x, we see differing patterns by context. In NYC, for instance, $r_x^{\text{SamePrin}} - r_x^{\text{DiffPrin}}$ is essentially zero beyond 5 years, while in OR it becomes larger as x increases. Beyond decreased sample size, sample selection is an important consideration in interpreting $r_x^{\text{SamePrin}} - r_x^{\text{DiffPrin}}$ at higher lags, as the r_x^{SamePrin} sample reflects a small minority of principals (particularly in OR) who remain in their school for more than five years.[19]

The results in Figure 2 suggest that the dynamic component of school performance is largely not driven by principal transitions. Instead, there exists within-school variation in student achievement performance that is semi-persistent and, thus, becomes erroneously attributed to principals. Comparing our variance estimates based on $r_x^{\text{SamePrin}} - r_x^{\text{DiffPrin}}$ to the baseline estimates from Table 1, only a small fraction—19% in math and 5% in reading—of the within-school variation in mean student test score performance used to produce principal VA estimates is explained by persistent effectiveness differences between principals. The remainder is likely the result of school factors that principals do not control.

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.[20] We do not argue that these rating scores are intended to measure the same construct 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,

(2020) incorporates the potential for drift in estimating principal VA, his residualization models include both principal and school fixed effects. Low mobility of principals and schools introduces substantial estimation error into the school and principal fixed effects (Bartanen and Husain 2022), which results in an overstatement of both the magnitude and inter-temporal stability of principal VA. We circumvent that issue here by only including a principal-by-school FE in the residualization step.

19. Section 5.4.3 presents suggestive evidence that the OR results at higher lags are driven by selection rather than principal effects. Using a matched sample approach that aims to address the selection issue, our pooled estimate of principal effects in OR decreases from 0.048 to 0.027 *SD* in math and 0.018 to 0.013 *SD* in reading.

20. 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.

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.

An additional way to demonstrate our key result is to implement a quasi-falsification test that estimates the "effects" of principals based on the observed history of principal switches for a different school. In essence, we simply reassign to each school the principal assignments from a different, randomly selected school. We then run our test-score variance decomposition models using these simulated principal assignments. Appendix Table A.5 shows these results for math across 100 iterations in each dataset, compared with the baseline estimates from Table 1 (Appendix Table A.6 shows the corresponding results in reading). Consistent with the small differences between r_x^{SamePrin} and r_x^{DiffPrin}, we find that the principal variance component estimates from the actual data are only slightly larger than those from the imputed data (between 0.001 and 0.019 *SD* units, depending on the context, subject and model). This result further reinforces that the observed between-principal variation in school performance is mostly capturing the effects of other school-level factors rather than the effect of the current principal.[21]

5.3 Event Studies

The results from Section 5.2 suggest that most of the variation used to form principal VA estimates reflects time-varying factors that would have occurred regardless of which principal was leading the school. As a further illustration of this finding, we shift to an event study framework that examines changes in school performance following the entry or exit of a high-estimated-VA (or low-estimated-VA) principal. This framework is conceptually similar

21. As further illustration, we show in Appendix G that once we explicitly model the autocorrelation of the error term in Equation 8, the principal-level variance components are effectively zero (Appendix Tables G.1 and G.2).

to Chetty, Friedman, and Rockoff (2014a), who examine changes in grade-level performance following the entry or exit of high- and low-VA teachers. Here, however, the logic is flipped; we do not expect to see a substantial change in school performance following the entry or exit of a high-estimated-VA principal insofar as principal VA mostly reflects factors unrelated to true principal effectiveness.

To implement this analysis, we first estimate principal VA in a model with principal-by-school and school fixed effects:

$$Y_{ispt} = \beta \mathbf{X}_{ist} + \gamma_s + \delta_{s,p} + \epsilon_{ispt} \tag{9}$$

where the control vector $\mathbf{X}_{ist}$ includes prior-year test scores and the full set of student and school characteristics (i.e., residualization model 2). We place a sum-to-zero constraint on the principal-by-school fixed effects $\delta_{s,p}$ within each school, such that $\hat{\delta}_{s,p}$ represents principal p's estimated performance in school s relative to the mean performance of all principals who ever led school s (Bartanen and Husain 2022). For the purposes of our event study, however, we cannot use this measure to identify high- and low-VA principals because the same student test score residuals contribute to both the school performance measure and the VA estimate, which will create a mechanical correlation (Chetty, Friedman, and Rockoff 2014a). Instead, we construct a leave-school-out VA measure $\hat{\delta}_{s,p}^{-s}$ using a precision-weighted average of $\hat{\delta}_{s,p}$ from prior and future schools that principal p leads.[22] We define high- and low-estimated-

22. This differs from Chetty, Friedman, and Rockoff (2014a) in that our jackknife measure leaves out all residuals from the principal's current school, including those that are outside of the event window. This is important because of the autocorrelation of the residuals. For instance, if year t is within the event window for examining changes in school performance and year $t+1$ is not, using year $t+1$ to form the principal's VA estimate will still create a mechanical correlation because many of the school-level factors (including those that principals do not control) in year t are also present in year $t+1$, and most of the students who contribute to year t performance also contribute to year $t+1$ performance (since most students will remain in the same school unless making a structural move). Notably, Cullen et al. (2021) propose a validation test of principal VA that is similar to ours, except that their principal VA estimates include residuals from the same school but outside of the event window (and not residuals from principals' tenures in other schools). They find that changes in principal VA positively predict (albeit weakly) changes in school performance, which they interpret as "evidence to suggest that our spell value-added estimates capture meaningful differences in principal effectiveness" (p.17). Our analysis instead suggests that this positive relationship is driven by the autocorrelation issues described above.

VA principals as the top and bottom quartiles using this leave-school-out measure. We then identify all events where a high- or low-estimated-VA principal enters or leaves a school, subject to the constraint that this principal leads the school for at least three years. For each event, we construct a six-year panel defined by event time -3 to 2, where 0 denotes the first year of the incoming principal. With this sample, we regress $\bar{Y}^*_{spt}$ on a set of indicator variables for event time:

$$\bar{Y}^*_{spt} = \sum_{k=-3}^{2} \beta_k \tau_k + \varphi_t + \varepsilon_{spt} \tag{10}$$

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 correlation of errors over time within each unique event.

Figure 4 demonstrates that there is no clear change in student test score performance following the arrival or departure of a high- or low-estimated-VA principal, on average. We plot the predicted margins from Equation 10 along with "first stage" results where the dependent variable is the leave-school-out VA estimate for the principal leading the school in time k.[23] Because NYC has relatively few principals who led multiple schools, event study results are less precise and, thus, we focus on the results from TN and OR.[24] To the extent that estimated principal VA reflects real differences in principals' causal effects on test scores, we should expect the change in VA to correspond to a change in student achievement. Despite a large change in estimated principal VA following each type of transition (based on the difference between the pooled pre- and post-years), nearly all of the estimated changes in student test scores are modest in magnitude and not statistically significant.

23. Note that to maximize our sample, we do not restrict to events where both the departing and incoming principal have a non-missing leave-school-out VA estimate (only the principal used to define the high- or low-VA entry/exit). As such, these "first stage" results include events where one side of the transition is dropped due to missing data. Restricting to events that have VA estimates for both principals yields very similar results, but they are less precise.

24. Results from NYC are shown in Appendix Figures A.6 and A.7. While less precise, these results are very similar to those from TN and OR. However, we cannot present the "first stage" results using value-added as the outcome because there are insufficient events where we can estimate a leave-school-out principal VA measure for *both* the departing and entering principal.

For example, the top left panel of Figure 4 shows results for the entry of a high-estimated-VA principal in Tennessee. This transition corresponds to, on average, a 0.15 *SD* increase in principal VA ($p < 0.001$) but a -0.005 *SD* increase in student test scores ($p = 0.92$). The one exception to this pattern of null results is the entry of a high-estimated-VA principal in Oregon. Here, we observe a 0.12 *SD* increase in principal VA and a 0.036 *SD* increase in math scores ($p = 0.02$). For reading achievement, we find null results across both contexts (see Figure 5). Standard errors for the simple pre-post difference in test scores range from 0.013 to 0.019, indicating that we cannot rule out modestly sized effects, including those of the magnitude from Table 2. On the whole, however, these results support the conclusion based on $r_x^{\text{SamePrin}} - r_x^{\text{DiffPrin}}$ that the true magnitude of principal effects is substantially smaller than previously reported and that most of the information contained in principal VA does not reflect persistent differences in principal quality.

5.4 Checking Threats to Validity

Fundamentally, our proposed framework rests on the idea that principal effects exhibit some persistence—either fixed or evolving through time. Empirically, we rely on our sample data to estimate $r_x^{\text{SamePrin}} - r_x^{\text{DiffPrin}}$, from which we can obtain the estimated magnitude of principal effects in terms of improved student outcomes. Under this framework, there are several potential threats to validity, including sorting bias, violation of the stationarity assumption, and sample selection bias. Below, we outline each of these threats and present evidence to demonstrate they are not likely to be major drivers of our results.

5.4.1 Sorting Bias

As described by Equation 6, the potential for non-random sorting of principals to schools means that the magnitude of principal effects based on $r_x^{\text{SamePrin}} - r_x^{\text{DiffPrin}}$ is a lower-bound estimate. Intuitively, if schools hire principals of similar quality, the stability of school

performance across principals (r_x^{DiffPrin}) will be greater, reducing the estimated magnitude of principal effects. Following Equation 6, we can conduct a bounding exercise based on assumptions about ρ_δ (the school-level intraclass correlation of principal quality).

As a first step, we obtain plausible values for ρ_δ by estimating its analog for observable principal characteristics, including years of principal experience (at the time of hire), years of assistant principal experience, advanced degree attainment, ratings from teachers and supervisors, and the principal's own value-added when they were a classroom teacher.[25] While there is no guarantee that any of these measures is highly correlated with principal value-added, they are characteristics available to school district administrators when determining principal hiring and placement. Appendix Table C.1 shows that, across nearly all of these measures, the estimate of ρ_δ is modest, ranging from 0 to 0.2.[26] Further, most newly hired principals have no prior principal experience, which may limit sorting insofar as principal quality is difficult to predict prior to observing someone in the role.

Appendix Table C.2 shows bias-corrected estimates of the magnitude of principal effects under different sorting scenarios. We average the pooled *SD* from Tables 2 and 3 across each of the three contexts. These estimates are 0.034 and 0.011 *SD* for math and reading VA, respectively. Applying Equation 6, we compute the bias-corrected estimate based on ρ_δ.[27]

25. Note that we do not have all of these measures in each context, either because they are unavailable or the sample size is too small. For example, we can estimate teacher value-added in Oregon only for the last six years of our panel, which means that very few (roughly 100) principals in our sample have teacher value-added estimates and, further, we do not observe in the same school multiple principals with teacher VA estimates, meaning that we cannot estimate the ICC for this measure.

26. For supervisor ratings in TN and teacher ratings in NYC, ρ_δ is roughly 0.3. However, part of the correlation in these perception measures likely reflects overlap in raters (i.e., many of the same supervisors or teachers are judging different principals within the same school, such that any rater-level variance will be incorporated into ρ_δ) and school-level factors that are incorrectly attributed by raters into principals' scores. Accordingly, we construct "leave-current-school-out" measures that average over a principal's scores when they led a different school, which should help to remove upward bias from overlap in raters (though only partially for supervisor ratings, because principals typically move within the same district and may have the same supervisor as their prior position). The ICC is attenuated, as expected, putting them more in line with results from the non-ratings measures.

27. Because we do not separately estimate $\sigma^2_{\delta^F}$ and $\sigma_{\delta_x^D}$ (the stable vs. dynamic components of principal quality), we need to make an additional assumption about their relative importance ($\frac{\sigma^2_{\delta^F}}{\sigma^2_{\delta^F}+\sigma_{\delta_x^D}}$) to solve for the magnitude of bias. Here, we assume this ratio is 1 (principal quality is perfectly stable), which will yield a larger magnitude of bias. However, our results are not particularly sensitive to this choice. See Appendix Table C.2 for details.

Under moderate sorting ($\rho_\delta = 0.2$), our revised estimates are 0.038 and 0.012 SD. Even with substantial sorting ($\rho_\delta = 0.6$), bias-corrected estimates are only slightly larger at 0.054 and 0.017 SD. Based on this exercise, we believe that our lower-bound estimates are likely to be close to the true magnitude of principal effects.

5.4.2 Stationarity Assumption

An additional potential threat to the validity of $r_x^{\text{SamePrin}} - r_x^{\text{DiffPrin}}$ as a indication of the magnitude of principal effects 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 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 D.1–D.3.[28] 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 D.1–D.3). In New York City, the larger correlations are only

28. 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 D.1–D.3.

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.

5.4.3 Selection Bias

The logic of using $r_x^{\text{SamePrin}} - r_x^{\text{DiffPrin}}$ to infer the magnitude of principal effects is that r_x^{DiffPrin} serves as a counterfactual: how stable would school performance have been between year t and year $t+x$ if a school had changed principals instead of keeping the same principal? An important practical issue to consider is the sample of schools who contribute to estimating r_x^{SamePrin} and r_x^{DiffPrin}. Principals' tenures tend to be short, so fewer schools contribute to r_x^{SamePrin} as x increases (see N_x^{same} in Table 2). Accordingly, one potential concern is that for larger x the characteristics of the schools for estimating r_x^{DiffPrin} and r_x^{SamePrin} are quite different, such that r_x^{DiffPrin} provides an inaccurate counterfactual. For example, some schools may have characteristics that make school performance more stable over time (independent of who leads them) and make principals more likely to remain in their position. When x is large, such schools may be overrepresented in r_x^{SamePrin} and underrepresented in r_x^{DiffPrin}, which would lead to upward bias in the estimated magnitude of principal effects.

To examine this potential source of bias, Appendix Tables E.1–E.3 show mean characteristics for the r_x^{SamePrin} and r_x^{DiffPrin} samples. In general, these sets of school-by-year pairs are observably similar; on average, they serve students with similar demographics, prior achievement, and attendance rates. Teacher composition in terms of years experience and value-added are also very similar. In Oregon, however, we observe that as x increases, the r_x^{SamePrin} sample becomes increasingly higher-achieving and less impoverished. This trend corresponds to the divergence of r_x^{SamePrin} and r_x^{DiffPrin} in Tables 2 and 3, whereby the esti-

mated magnitude of principal effects was increasing with x. One potential explanation, then, is that this divergence is a product of increasing imbalance between the samples for r_x^{SamePrin} and r_x^{DiffPrin}, whereby r_x^{DiffPrin} is an increasingly poor counterfactual for the non-principal-driven stability of schools that keep their principal for an extended period.

As an additional check, we re-estimate r_x^{SamePrin} and r_x^{DiffPrin} in Oregon using a matched sample of schools that are more balanced on observables.[29] Specifically, we employ a coarsened exact matching approach using strata defined by schools' shares of low-income and white students, as well as students' prior-school achievement scores (see Appendix E for details on the matching procedure). Using the matched samples, we no longer observe that $r_x^{\text{SamePrin}} - r_x^{\text{DiffPrin}}$ increases with x, which provides suggestive evidence that the larger SD estimates in OR for higher lags are driven by selection rather than true principal effects.

5.5 Examining Mechanisms

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

29. We also implemented the same matching approach in Tennesee and New York City. However, the r_x^{SamePrin} and r_x^{DiffPrin} results were very similar to our baseline results, which is unsurprising given that r_x^{SamePrin} and r_x^{DiffPrin} schools were already observably similar.

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.[30] This avoids estimating teacher VA from the same test score residuals that form our school performance measures. Figure 6 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 $\bar{Y}^*_{spt}$.[31] To the extent that principals can meaningfully shape

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

31. Appendix Tables A.7–A.9 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

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).[32]

The second mechanism we investigate is the composition of entering cohorts of students 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 7 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. 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 sug-

performance measure.

32. 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.

gestive 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.

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. 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, 46 states now incorporate school-level, student-outcome-based measures into principal evaluation systems (Donaldson et al. 2021). 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 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 mostly 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 concerns the magnitude of principals' effects on student outcomes. Once accounting for the dynamic nature of school performance, variation in principal quality explains relatively little of the observed variation

in student test scores or attendance. Our estimates imply that a 1 *SD* increase in principal quality raises student test scores by 0.03 *SD* in math and 0.01 *SD* in reading. These magnitudes refute the existing literature, where principals' estimated impacts are much larger in magnitude. An important exception is Laing et al. (2016), whose variance-based estimates using similar methodology to ours yield somewhat smaller estimates than found in the rest of the literature. Our estimates are still consistent with the notion that principals can play a role in improving student outcomes. Despite relatively small magnitudes at the individual student level, principals' impacts apply to hundreds of students across an entire school.

The second finding is that, at least as currently implemented, attempts to identify high-quality principals through student outcomes are fundamentally 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 and unreliable estimates of individual principals' 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. While districts or states do not currently employ principal value-added models, they often rely on even cruder test-score-based measures, such as school value-added or changes in proficiency rates, to formally or informally evaluate principals (Donaldson et al. 2021). These measures likely suffer from the same misattribution and imprecision issues we demonstrate here.

Given that many principals do not remain in the same school beyond a few years, we suggest some caution in the interpretation of our findings concerning the magnitude of principal effects. The frequent churn of school leaders could be part of the substantive explanation for why we observe little variation in principals' measured impacts. 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.

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

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, and Aliza N Husain. 2022. "Connected networks in principal value-added models." *Economics of Education Review* 90:102292.

Blackwell, Matthew, Stefano Iacus, Gary King, and Giuseppe Porro. 2009. "cem: Coarsened exact matching in Stata." *The Stata Journal* 9 (4): 524–546.

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.

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.

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.

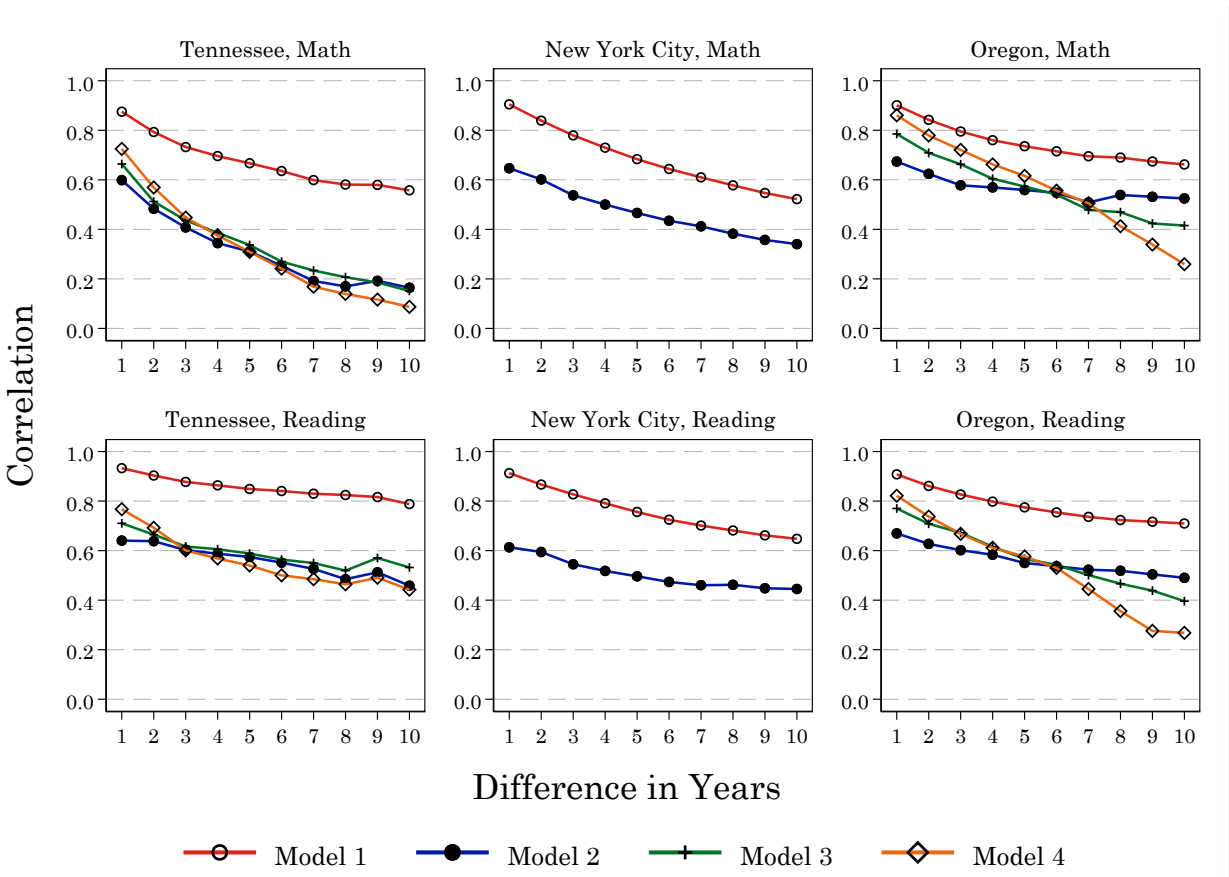


Figure 1: Autocorrelation Vectors

Notes: Figures report autocorrelation "drift" vectors generated from sample-size-precision-weighted correlations of school-by-year mean residualized test scores generated from Equation 7. 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 Tables 2 and 3.

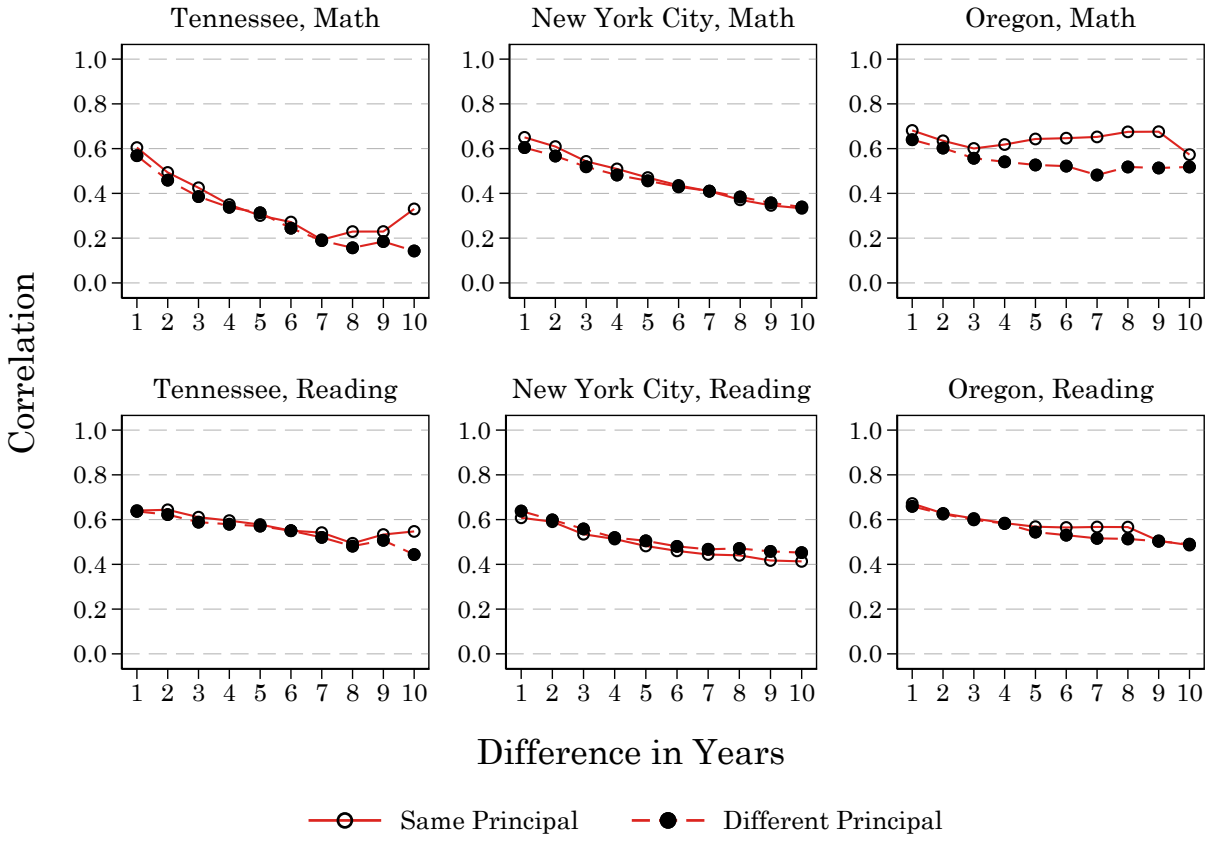
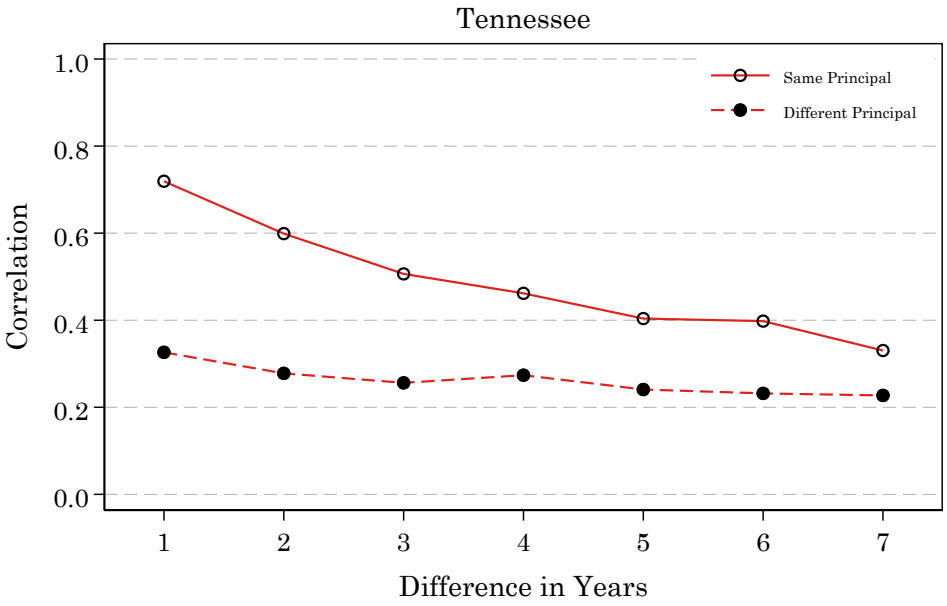


Figure 2: Autocorrelations Within and Between Principals

Notes: Figures report autocorrelation "drift" vectors generated from sample-size-precision-weighted correlations of school-by-year mean residualized test scores generated from Equation 7. 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 Tables 2 and Table 3.

(a) Supervisor Ratings



(b) Teacher Ratings

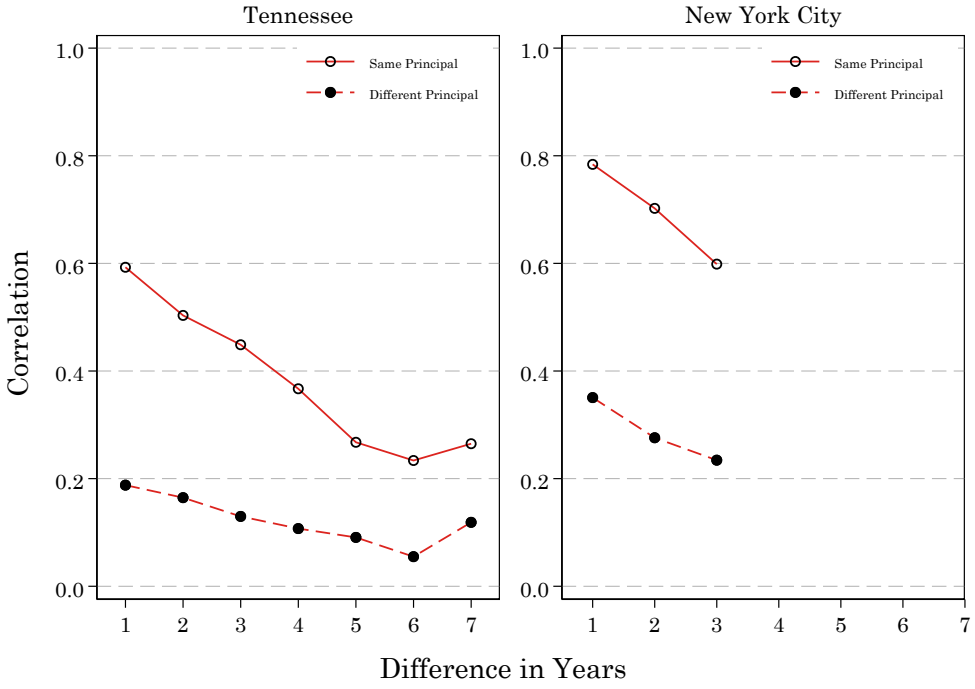


Figure 3: Autocorrelations of Perceptions of Performance, Within and Between Principals

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.

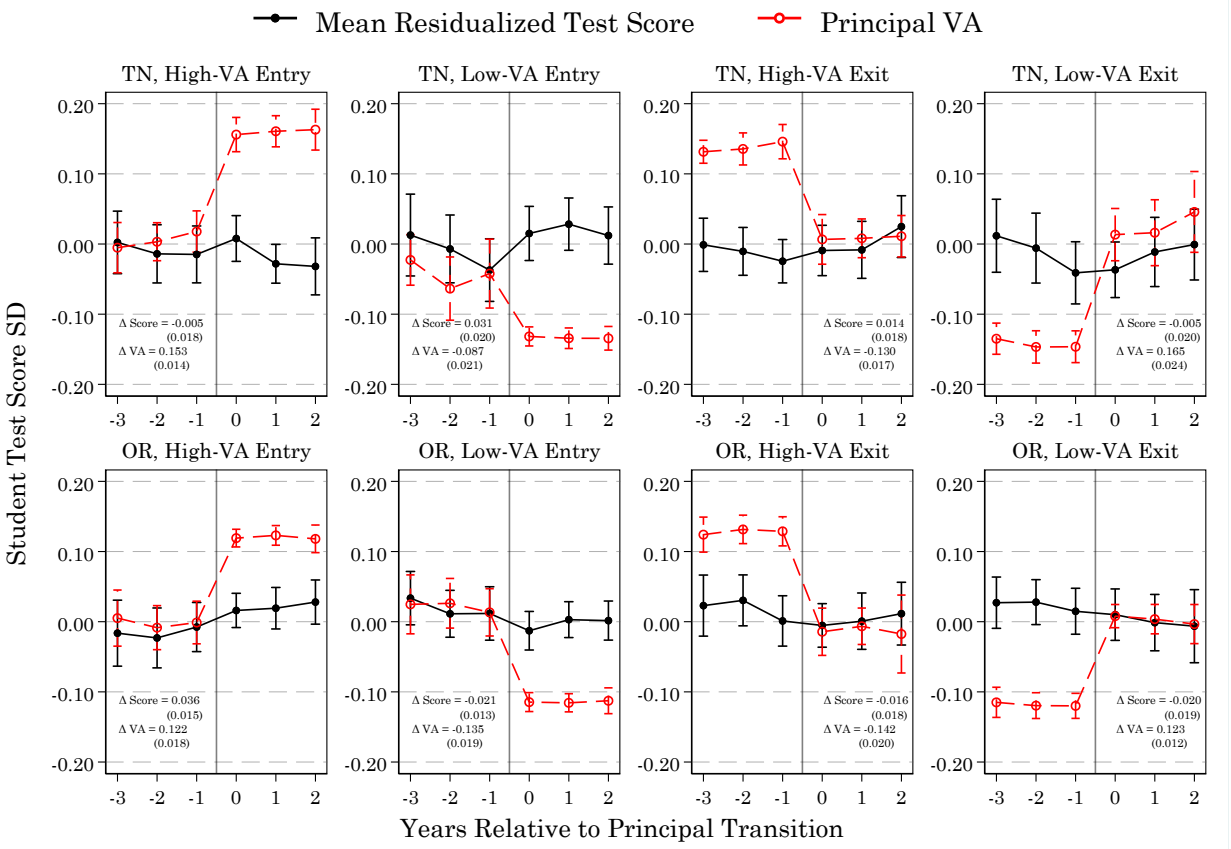


Figure 4: Event Study (Math)

Notes: Figures report event-study estimates and 95% confidence intervals from Equation 10. Residualization models adjust for student demographic characteristics and cubic polynomials of lagged-test scores and attendance (Model 2). High-performing and low-performing principals are defined by current-school-leave-out principal VA measure generated from Equation 9. 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.

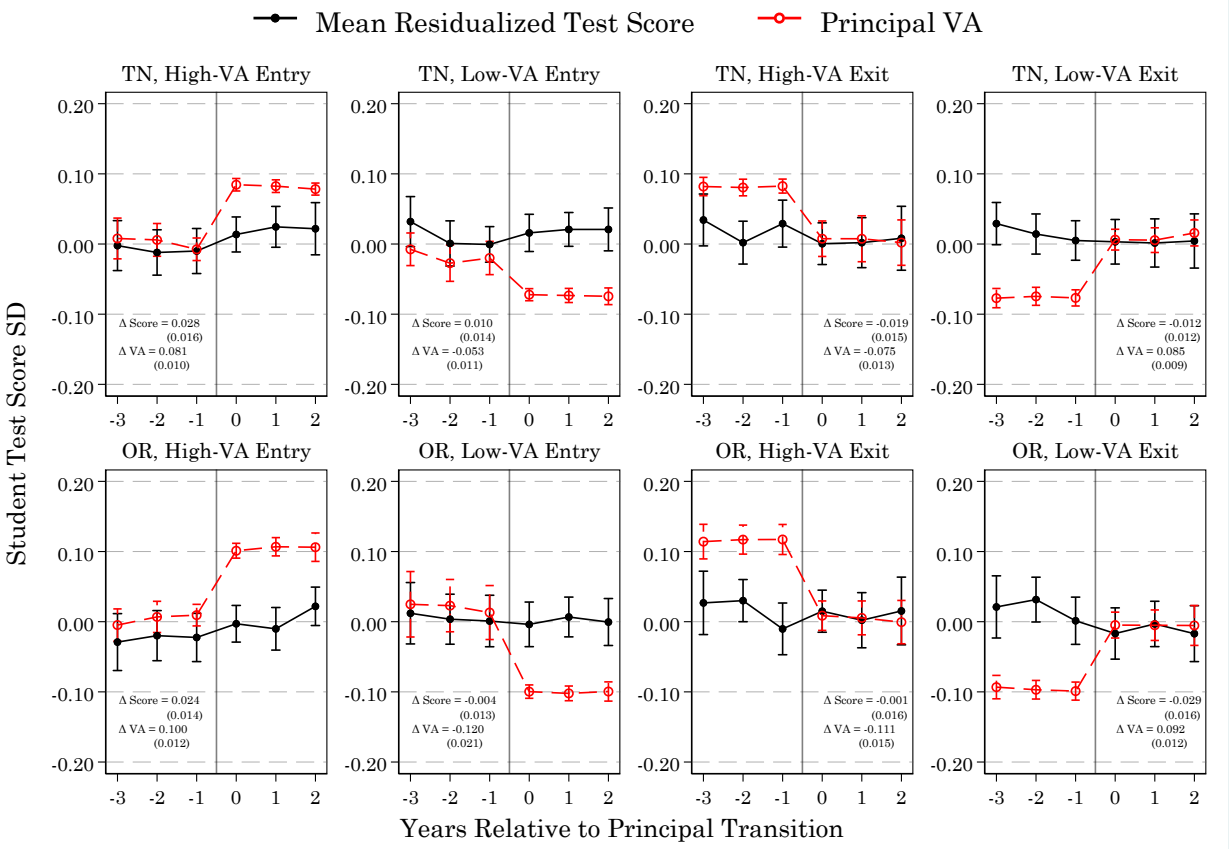


Figure 5: Event Study (Reading)

Notes: Figures report event-study estimates and 95% confidence intervals from Equation 10. Residualization models adjust for student demographic characteristics and cubic polynomials of lagged-test scores and attendance (Model 2). High-performing and low-performing principals defined by current-school-leave-out principal VA measure generated from Equation 9. 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.

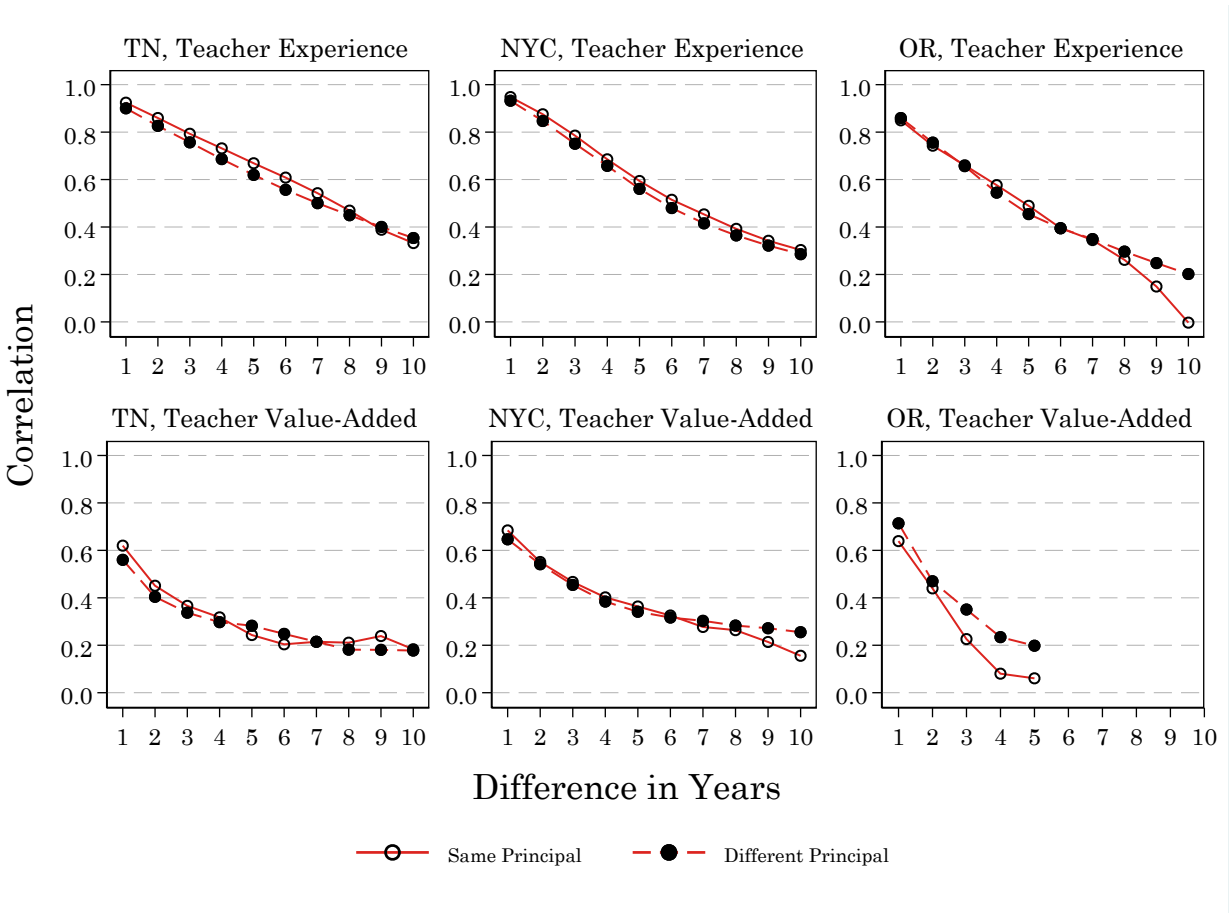


Figure 6: Autocorrelation Vectors for Teacher Composition

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.

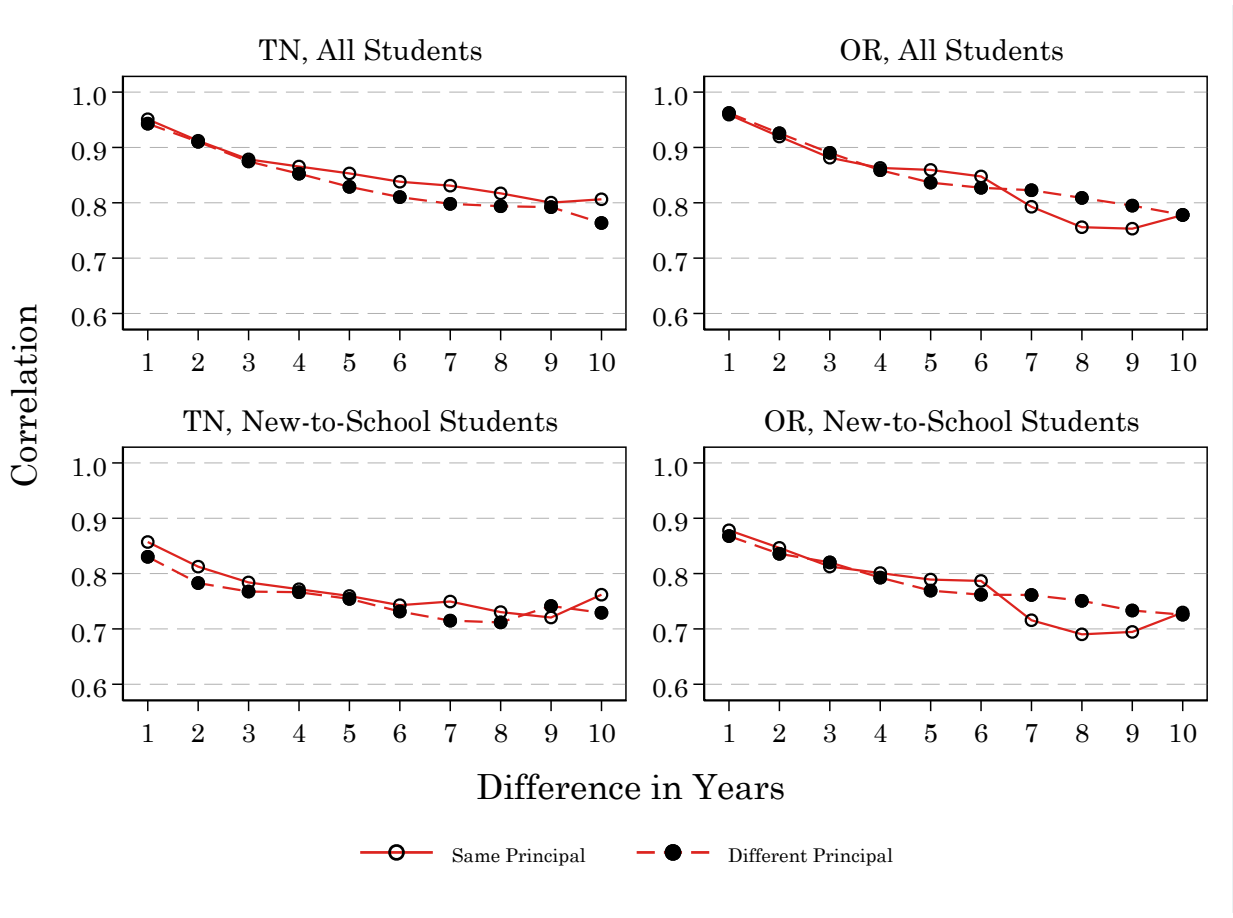


Figure 7: Autocorrelation Vectors for Student Composition Using Prior-School Lagged Test Scores

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 7. 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 9^{th}-grade students appear in our sample prior to 2014 and none afterwards.We cannot produce these results for NYC, as we do not have information about when a student first enrolls in a school.

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	0.128	0.173	0.161
Principal	0.136	0.136	0.094	0.111	0.121
Residual	0.173	0.168	0.154	0.162	0.147
Variance Components (%)					
School	77.7	65.1	33.2	43.8	41.7
Principal	8.5	13.9	18.2	17.9	23.5
Residual	13.8	21.1	48.6	38.3	34.7
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	0.236	0.114		
Principal	0.170	0.129	0.061		
Residual	0.150	0.138	0.112		
Variance Components (%)					
School	83.0	61.0	44.2		
Principal	9.6	18.2	12.8		
Residual	7.5	20.8	43.0		
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	0.322	0.169	0.243	0.265
Principal	0.117	0.114	0.074	0.100	0.096
Residual	0.143	0.137	0.138	0.146	0.127
Variance Components (%)					
School	79.5	76.5	54.1	65.5	73.4
Principal	8.2	9.6	10.2	11.0	9.7
Residual	12.3	13.9	35.7	23.5	16.9
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 8. 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: Testing Differences in Autocorrelations (Math)

x	N_x^{same}	N_x^{diff}	r_x^{same}	r_x^{diff}	diff_x	p-val	SD_x
Panel A: Tennessee							
1 year	11846	2460	0.605	0.569	0.036	0.012	0.035
2 years	8429	4023	0.494	0.459	0.034	0.020	0.034
3 years	6038	4928	0.425	0.386	0.040	0.014	0.036
4 years	4670	5853	0.350	0.338	0.012	0.497	0.020
5 years	3181	5758	0.301	0.313	-0.012	0.539	0.000
6 years	2125	5254	0.272	0.244	0.028	0.250	0.030
7 years	1312	4545	0.193	0.189	0.004	0.895	0.012
8 years	786	3589	0.229	0.157	0.072	0.056	0.049
9 years	599	3416	0.230	0.185	0.045	0.292	0.039
10 years	340	2426	0.331	0.143	0.188	0.001	0.079
Pooled *SD* Estimate							0.032
Panel B: New York City							
1 year	14848	2015	0.651	0.605	0.046	0.001	0.030
2 years	11900	3693	0.609	0.567	0.042	0.001	0.029
3 years	9437	4924	0.543	0.519	0.024	0.056	0.022
4 years	7462	5728	0.509	0.482	0.027	0.039	0.023
5 years	5897	6136	0.471	0.456	0.014	0.314	0.017
6 years	4629	6278	0.436	0.430	0.006	0.696	0.011
7 years	3597	6201	0.410	0.410	-0.001	0.975	0.000
8 years	2745	5974	0.371	0.384	-0.013	0.518	0.000
9 years	2019	5656	0.346	0.358	-0.011	0.617	0.000
10 years	1428	5233	0.333	0.339	-0.006	0.823	0.000
Pooled *SD* Estimate							0.022
Panel C: Oregon							
1 year	8066	2004	0.681	0.640	0.041	0.003	0.041
2 years	5543	3374	0.635	0.603	0.032	0.016	0.036
3 years	3675	4180	0.601	0.557	0.044	0.004	0.042
4 years	2374	4506	0.619	0.541	0.077	0.000	0.056
5 years	1470	4475	0.643	0.527	0.116	0.000	0.068
6 years	894	4162	0.647	0.522	0.125	0.000	0.071
7 years	534	3678	0.653	0.482	0.170	0.000	0.082
8 years	303	3040	0.675	0.518	0.157	0.000	0.079
9 years	171	2320	0.676	0.514	0.162	0.001	0.080
10 years	86	1559	0.573	0.518	0.055	0.486	0.047
Pooled *SD* Estimate							0.048

Notes: Table reports r^{same} and r^{diff} from sample-size-precision-weighted correlations of school-by-year mean residualized test score generated from Equation 7. Correlations are between year t and $t+x$ for the same school. r^{same} denotes the sub-sample of school-by-year pairs where the principal is the same in both years. r^{diff} 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). diff shows the difference in the correlations and p-val is the two-tailed p-value for the null hypothesis that the difference in correlations is zero. *SD* reports the estimated standard deviation of principal value-added based on Equation 5, which multiplies the difference in correlations by the variance of school-by-year mean residuals. When the difference in correlations is negative, we report an estimated *SD* of zero. The bottom row of each panel shows the pooled variance (reported as SD) estimate, which is obtained by computing the average of diff_x (weighted by N_x^{same}) and multiplying by σ_Y^2.

Table 3: Testing Differences in Autocorrelations (Reading)

x	N_x^{same}	N_x^{diff}	r_x^{same}	r_x^{diff}	diff_x	p-val	SD_x
Panel A: Tennessee							
1 year	11880	2468	0.640	0.637	0.004	0.784	0.008
2 years	8457	4038	0.644	0.622	0.021	0.062	0.019
3 years	6056	4947	0.611	0.589	0.022	0.068	0.020
4 years	4685	5875	0.595	0.579	0.016	0.206	0.017
5 years	3189	5775	0.578	0.571	0.007	0.639	0.011
6 years	2131	5272	0.551	0.550	0.001	0.971	0.003
7 years	1313	4565	0.542	0.521	0.021	0.345	0.019
8 years	785	3599	0.494	0.481	0.013	0.668	0.015
9 years	598	3420	0.533	0.508	0.026	0.425	0.021
10 years	339	2426	0.547	0.444	0.103	0.018	0.043
Pooled *SD* Estimate							0.016
Panel B: New York City							
1 year	14832	2017	0.608	0.638	-0.030	0.038	0.000
2 years	11883	3693	0.591	0.599	-0.008	0.500	0.000
3 years	9422	4926	0.535	0.558	-0.024	0.054	0.000
4 years	7450	5728	0.513	0.520	-0.007	0.613	0.000
5 years	5888	6132	0.482	0.505	-0.023	0.096	0.000
6 years	4623	6272	0.460	0.480	-0.020	0.180	0.000
7 years	3591	6195	0.444	0.467	-0.023	0.163	0.000
8 years	2741	5966	0.440	0.471	-0.031	0.094	0.000
9 years	2016	5650	0.417	0.458	-0.041	0.052	0.000
10 years	1426	5226	0.413	0.453	-0.039	0.106	0.000
Pooled *SD* Estimate							0.000
Panel C: Oregon							
1 year	8065	2006	0.672	0.659	0.013	0.355	0.020
2 years	5544	3375	0.628	0.625	0.003	0.848	0.009
3 years	3676	4179	0.605	0.599	0.006	0.680	0.014
4 years	2374	4503	0.584	0.583	0.002	0.914	0.008
5 years	1469	4473	0.568	0.544	0.024	0.243	0.027
6 years	893	4161	0.565	0.531	0.034	0.184	0.033
7 years	535	3682	0.567	0.516	0.051	0.121	0.040
8 years	303	3041	0.567	0.514	0.053	0.218	0.041
9 years	171	2320	0.505	0.503	0.002	0.969	0.009
10 years	86	1560	0.486	0.490	-0.004	0.962	0.000
Pooled *SD* Estimate							0.018

Notes: Table reports r^{same} and r^{diff} from sample-size-precision-weighted correlations of school-by-year mean residualized test score generated from Equation 7. Correlations are between year t and $t+x$ for the same school. r^{same} denotes the sub-sample of school-by-year pairs where the principal is the same in both years. r^{diff} 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). diff shows the difference in the correlations and p-val is the two-tailed p-value for the null hypothesis that the difference in correlations is zero. *SD* reports the estimated standard deviation of principal value-added based on Equation 5, which multiplies the difference in correlations by the variance of school-by-year mean residuals. When the difference in correlations is negative, we report an estimated *SD* of zero. The bottom row of each panel shows the pooled variance (reported as SD) estimate, which is obtained by computing the average of diff_x (weighted by N_x^{same}) and multiplying by σ^2_Y.

A Supplemental Figures and Tables

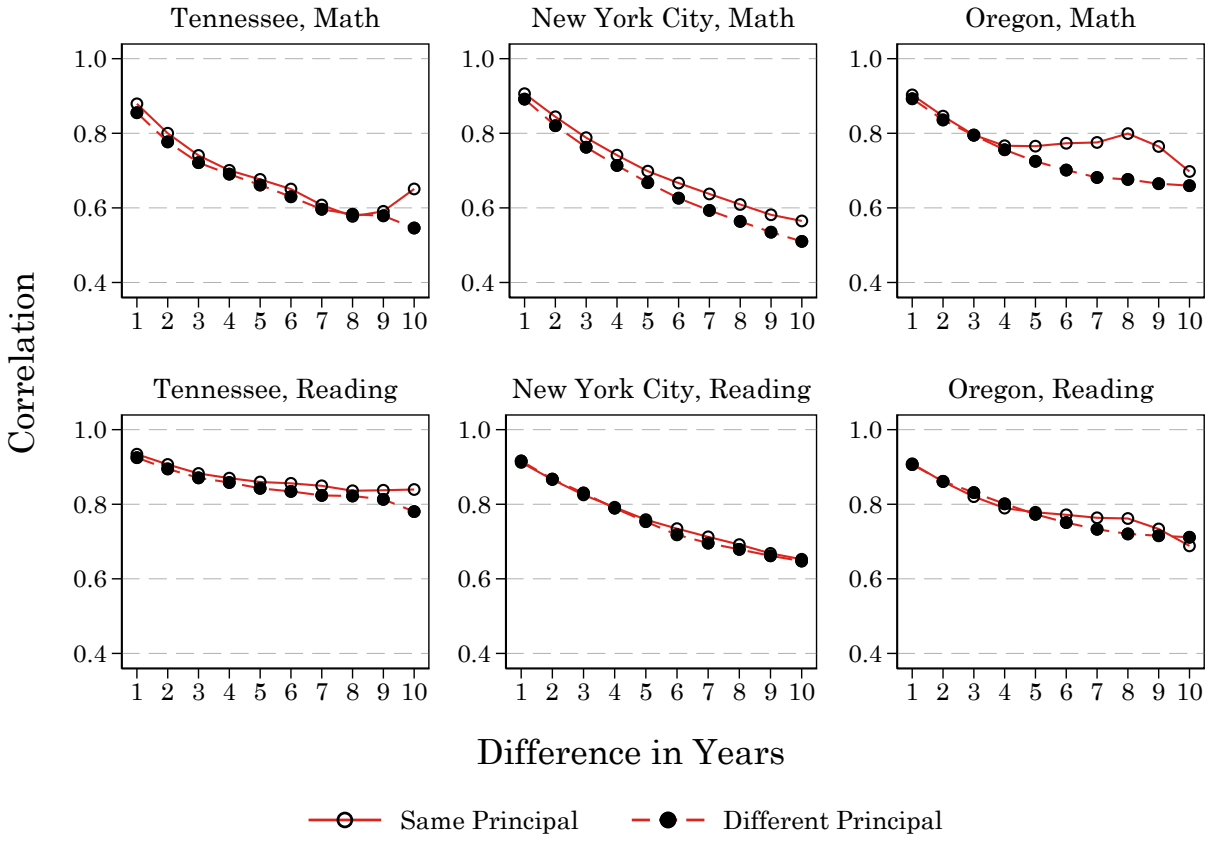


Figure A.1: Autocorrelations Within and Between Principals (Model 1)

Notes: Figures report autocorrelation "drift" vectors generated from sample-size-precision-weighted correlations of school-by-year mean residualized test score generated from Equation 7. 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 (Model 1).

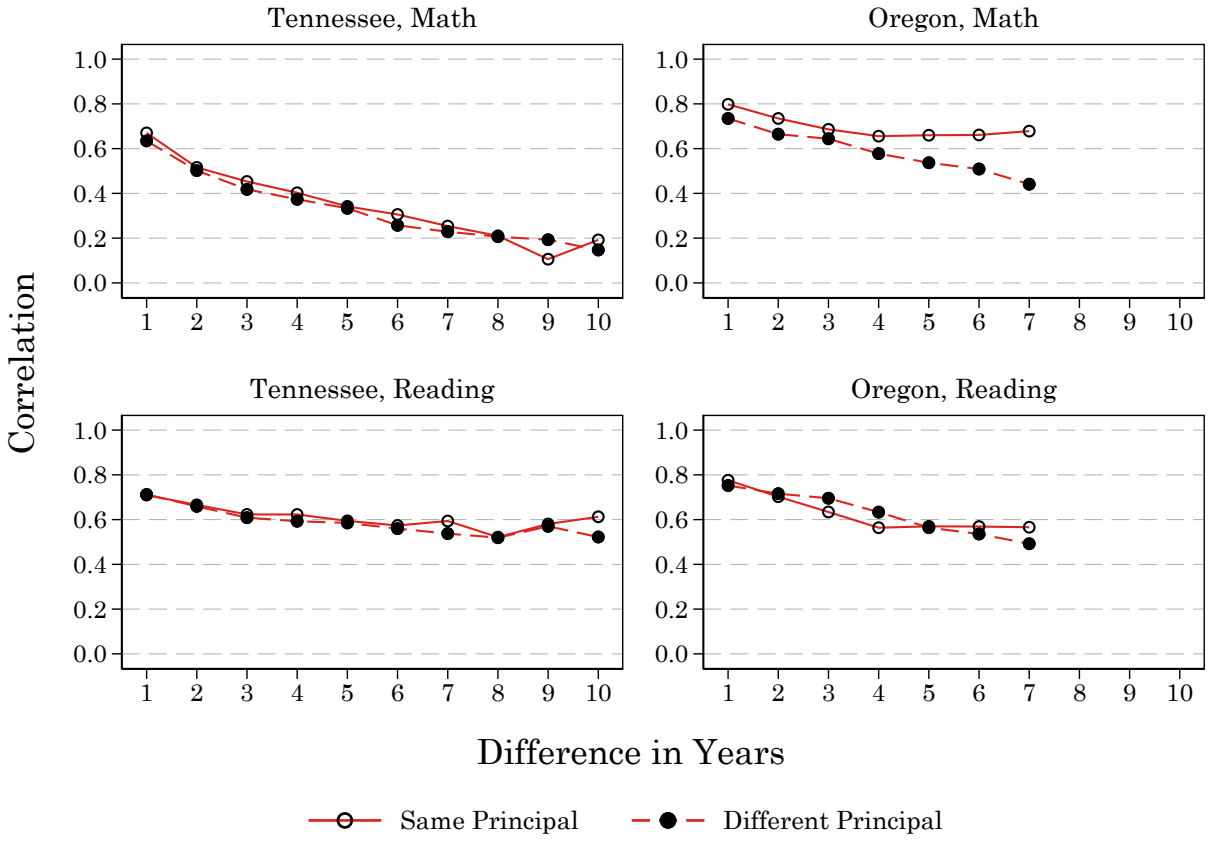


Figure A.2: Autocorrelations Within and Between Principals (Model 3)

Notes: Figures report autocorrelation "drift" vectors generated from sample-size-precision-weighted correlations of school-by-year mean residualized test score generated from Equation 7. 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 for new-to-school students only (Model 3). Model 3 not estimated for NYC because we do not observe year of first enrollment. Due to very small cell sizes (< 100), we do not report the correlations for 8 years or above in Oregon.

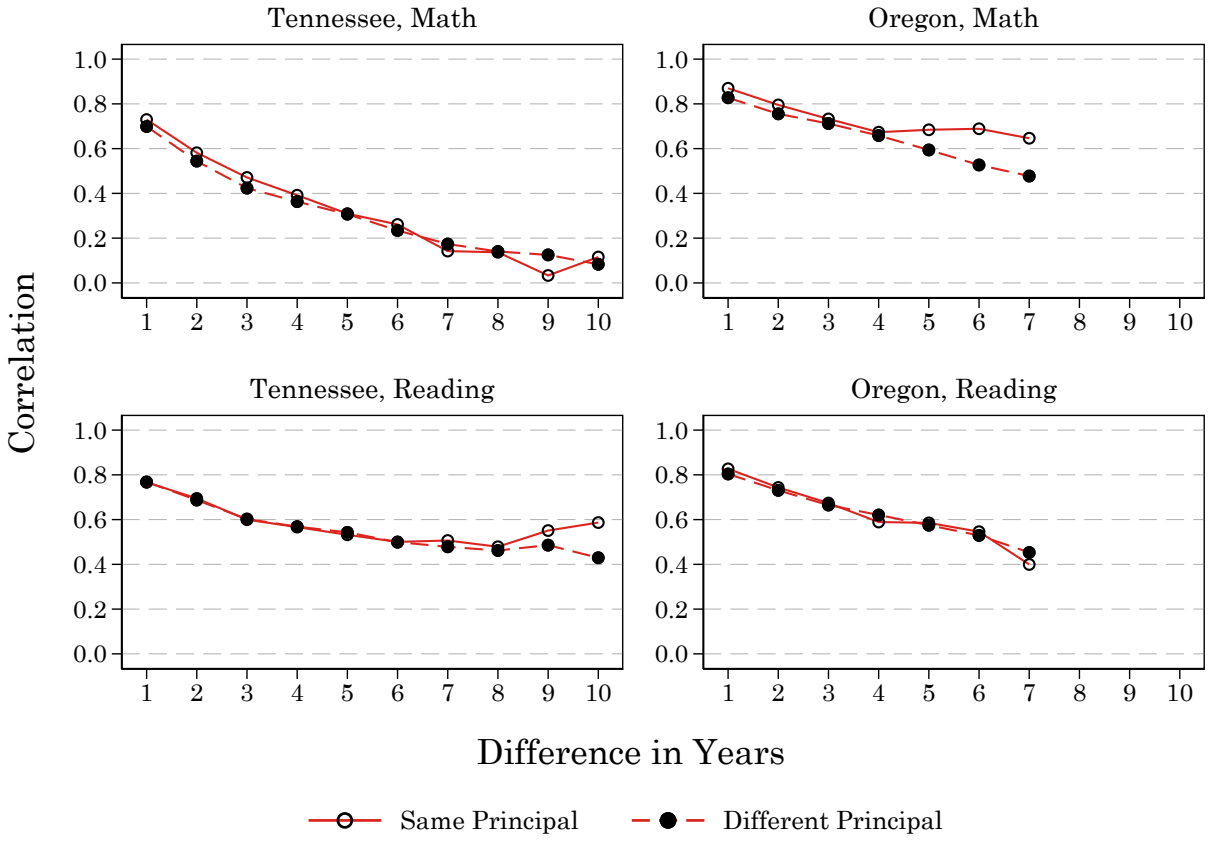
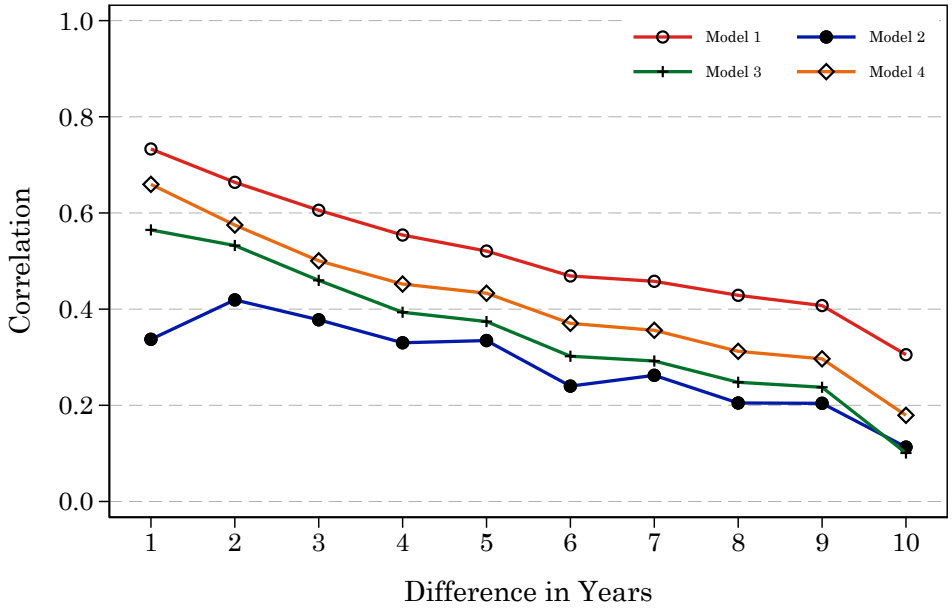


Figure A.3: Autocorrelations Within and Between Principals (Model 4)

Notes: Figures report autocorrelation "drift" vectors generated from sample-size-precision-weighted correlations of school-by-year mean residualized test score generated from Equation 7. 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 from prior-school (Model 4). Model 4 not estimated for NYC because we do not observe year of first enrollment. Due to very small cell sizes (< 100), we do not report the correlations for 8 years or above in Oregon.

(a) Within-School



(b) Within- and Between-Principals

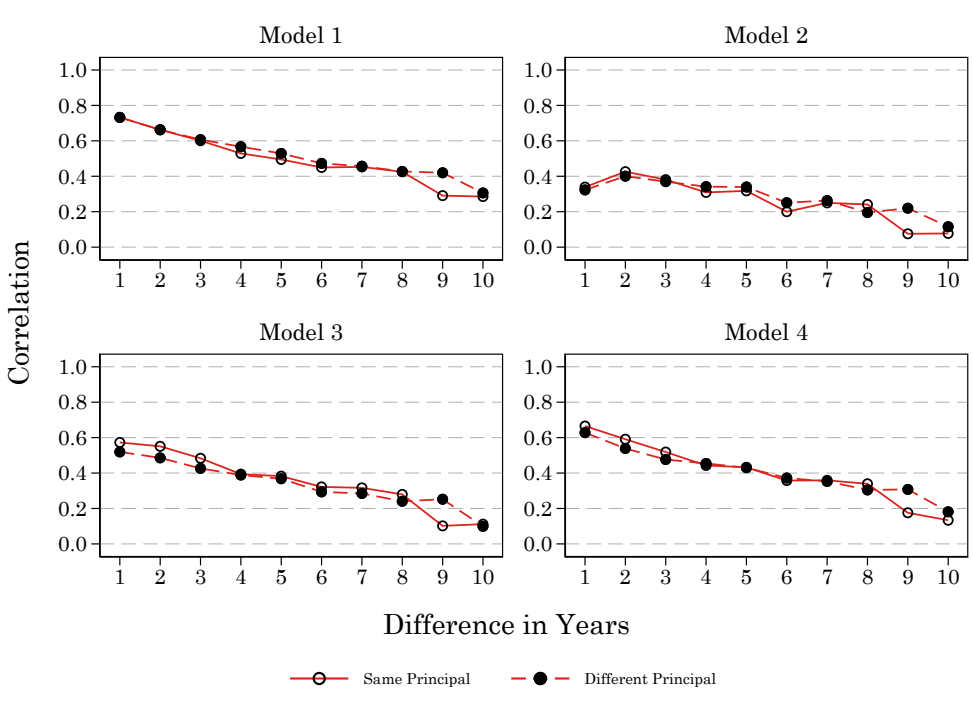
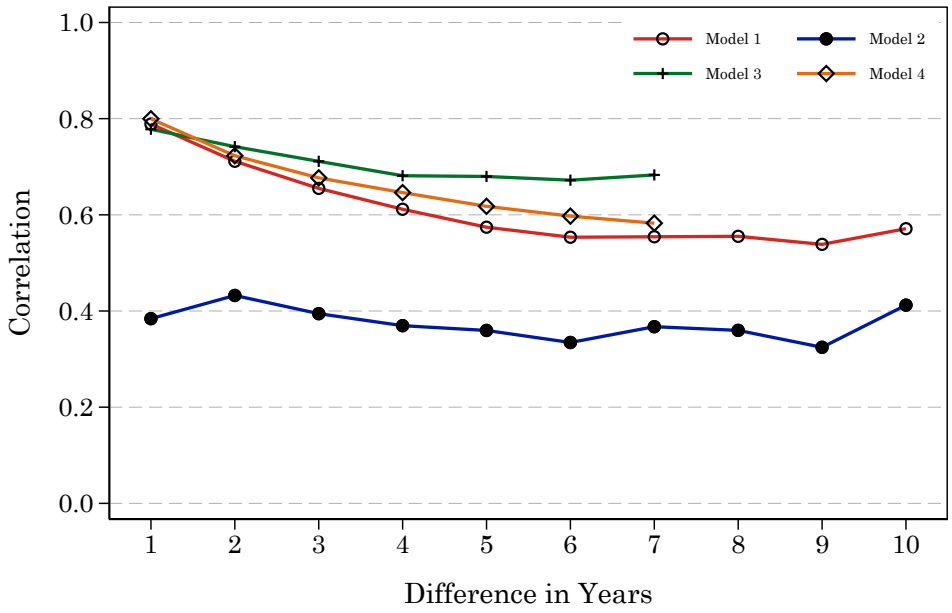


Figure A.4: Autocorrelation Vectors (TN Attendance)

Notes: Figures report autocorrelation "drift" vectors generated from sample-size-precision-weighted correlations of school-by-year mean residualized test score generated from Equation 7. 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

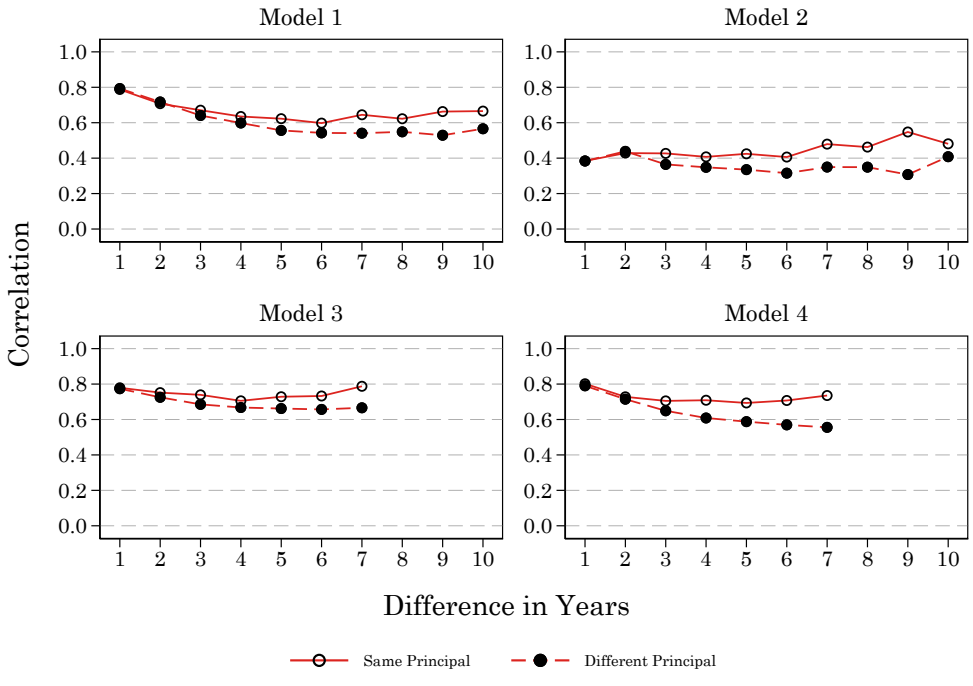


Figure A.5: Autocorrelation Vectors (OR Attendance)

Notes: Figures report autocorrelation "drift" vectors generated from sample-size-precision-weighted correlations of school-by-year mean residualized test score generated from Equation 7. 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.

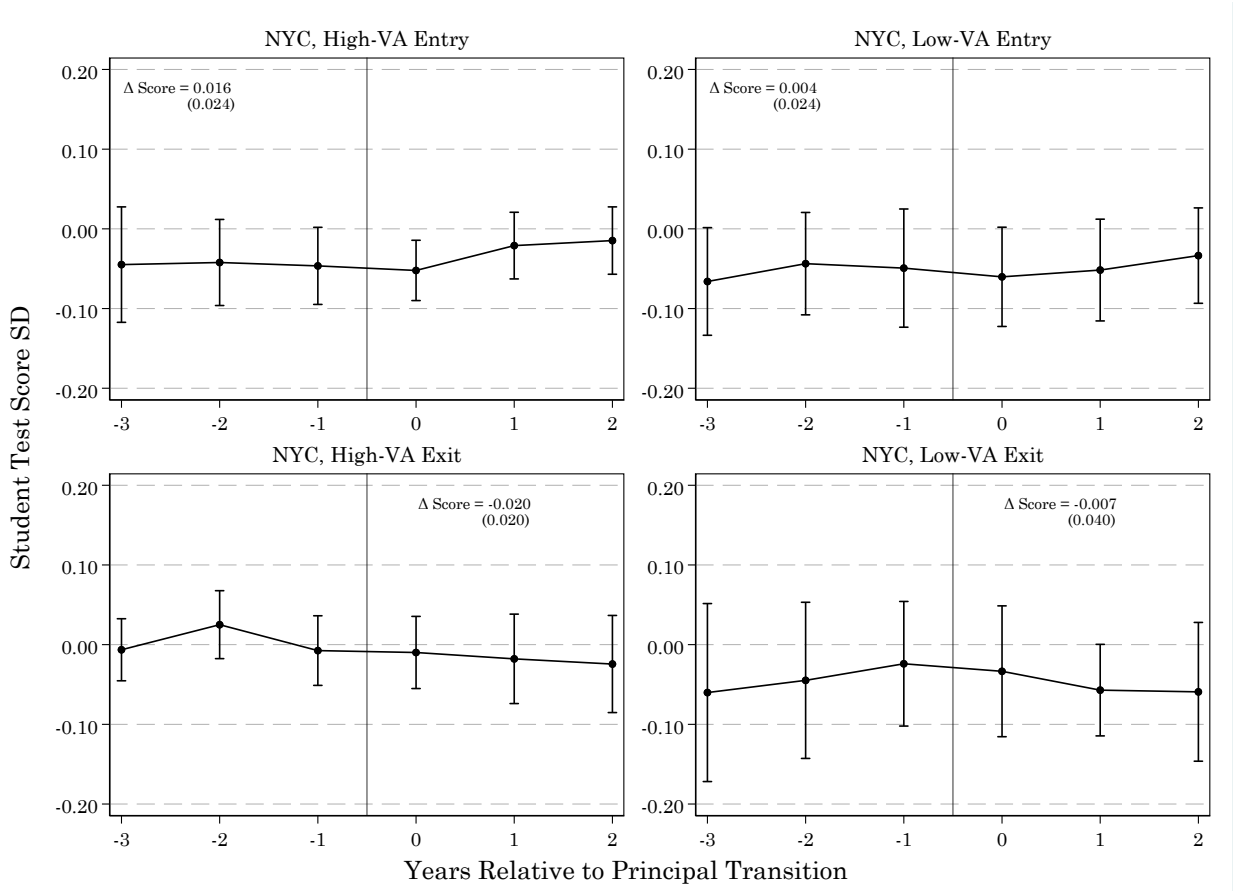


Figure A.6: Event Study (NYC, Math)

Notes: Figures report event-study estimates and 95% confidence intervals from Equation 10. Residualization models adjust for student demographic characteristics and cubic polynomials of lagged-test scores and attendance (Model 2). High-performing and low-performing principals defined by current-school-leave-out principal VA measure generated from Equation 9. We do not present "first stage" results using principal value-added as the outcome because there are insufficient events in NYC where we can estimate a leave-school-out principal VA measure for *both* the departing and entering principal. 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.

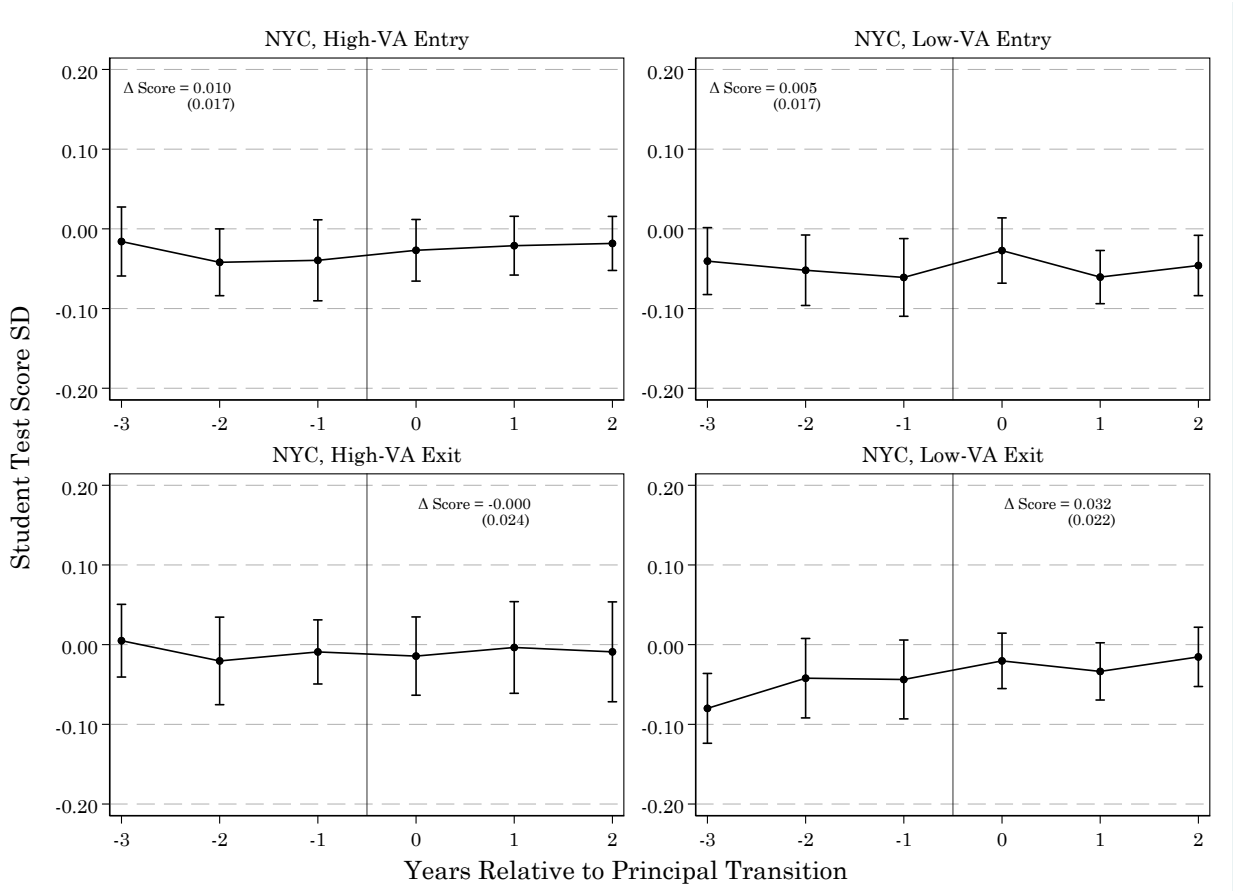


Figure A.7: Event Study (NYC, Reading)

Notes: Figures report event-study estimates and 95% confidence intervals from Equation 10. Residualization models adjust for student demographic characteristics and cubic polynomials of lagged-test scores and attendance (Model 2). High-performing and low-performing principals defined by current-school-leave-out principal VA measure generated from Equation 9. We do not present "first stage" results using principal value-added as the outcome because there are insufficient events in NYC where we can estimate a leave-school-out principal VA measure for *both* the departing and entering principal. 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.

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	0.294	0.119	0.167	0.143
Principal	0.092	0.085	0.045	0.056	0.065
Residual	0.119	0.113	0.099	0.113	0.096
Variance Components (%)					
School	89.2	81.2	54.5	63.7	60.5
Principal	4.0	6.8	7.9	7.2	12.6
Residual	6.8	11.9	37.6	29.2	26.9
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	0.239	0.110		
Principal	0.143	0.101	0.041		
Residual	0.135	0.122	0.101		
Variance Components (%)					
School	86.0	69.6	50.6		
Principal	7.4	12.5	6.9		
Residual	6.6	18.0	42.5		
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	0.289	0.152	0.219	0.202
Principal	0.097	0.092	0.060	0.081	0.075
Residual	0.126	0.120	0.124	0.129	0.110
Variance Components (%)					
School	82.0	78.4	54.9	67.5	69.6
Principal	6.6	8.0	8.6	9.2	9.7
Residual	11.4	13.6	36.5	23.3	20.7
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 8. 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 8. 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 8. 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: 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	0.012***	0.005***	0.002***	0.000	0.004***	0.000*	0.002***	-0.000*	0.010***	0.004***	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		0.012***		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		0.015***		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		0.017***		0.002***		0.007***		0.004***		0.009***		0.008***
		(0.001)		(0.000)		(0.000)		(0.000)		(0.001)		(0.001)
Diff = 6 years		0.020***		0.003***		0.008***		0.005***		0.011***		0.010***
		(0.001)		(0.000)		(0.000)		(0.000)		(0.001)		(0.001)
Diff = 7 years		0.024***		0.004***		0.009***		0.005***		0.015***		0.012***
		(0.001)		(0.001)		(0.000)		(0.000)		(0.001)		(0.001)
Diff = 8 years		0.025***		0.005***		0.009***		0.005***		0.014***		0.014***
		(0.001)		(0.001)		(0.001)		(0.000)		(0.002)		(0.001)
Diff = 9 years		0.021***		0.004***		0.010***		0.005***		0.017***		0.018***
		(0.001)		(0.001)		(0.001)		(0.000)		(0.002)		(0.001)
Diff = 10 years		0.018***		0.005***		0.011***		0.005***		0.025***		0.026***
		(0.002)		(0.001)		(0.001)		(0.000)		(0.002)		(0.002)
Diff = 11 years		0.016***		0.003***		0.010***		0.005***		0.046***		0.048***
		(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						0.010***		0.007***				
						(0.001)		(0.001)				
Diff = 15 years						0.009***		0.006***				
						(0.001)		(0.001)				
Diff = 16 years						0.010***		0.005***				
						(0.001)		(0.001)				
Diff = 17 years						0.008***		0.006***				
						(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	0.006	0.003	0.001	0.000	0.002	0.000	0.001	0.000	0.005	0.002	0.004	0.001
Within-school SD	0.079	0.052	0.030	0.014	0.044	0.015	0.028	0.000	0.071	0.043	0.064	0.030
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.5: Falsification Tests for Principal Variance Component (Math)

	Model 0	Model 1	Model 2	Model 3	Model 4
Panel A: Tennessee					
Actual Data	0.136	0.136	0.094	0.111	0.121
Imputed Data (mean of 100 iterations)	0.121	0.121	0.082	0.097	0.105
Imputed Data (min)	0.115	0.115	0.075	0.089	0.096
Imputed Data (max)	0.127	0.127	0.088	0.105	0.111
Panel B: New York City					
Actual Data	0.170	0.129	0.061		
Imputed Data (mean of 100 iterations)	0.151	0.120	0.056		
Imputed Data (min)	0.142	0.114	0.052		
Imputed Data (max)	0.160	0.126	0.060		
Panel C: Oregon					
Actual Data	0.117	0.114	0.074	0.100	0.096
Imputed Data (mean of 100 iterations)	0.108	0.105	0.064	0.085	0.086
Imputed Data (min)	0.101	0.098	0.056	0.073	0.076
Imputed Data (max)	0.116	0.112	0.072	0.097	0.100
Student Characteristics		✓	✓	✓	✓
Prior-Year Test Scores			✓	✓	
New-to-School Students Only				✓	
Prior-School Test Scores					✓

Notes: Cells report standard deviations of principal variance components estimated from Equation 8. Actual Data reports the estimate shown in Table 1. To run our falsification tests, we reassign each school to have a different principal assignment history based on the observed history from another school in the dataset (picked at random). We then estimate our variance decomposition models using these imputed principal assignments. We report the mean across 100 iterations for the estimated principal variance components (in *SD* units). We also report the minimum and maximum estimate across the 100 iterations.

Table A.6: Falsification Tests for Principal Variance Component (Reading)

	Model 0	Model 1	Model 2	Model 3	Model 4
Panel A: Tennessee					
Actual Data	0.092	0.085	0.045	0.056	0.065
Imputed Data (mean of 100 iterations)	0.083	0.078	0.039	0.050	0.059
Imputed Data (min)	0.078	0.073	0.035	0.043	0.053
Imputed Data (max)	0.088	0.082	0.044	0.060	0.066
Panel B: New York City					
Actual Data	0.143	0.101	0.041		
Imputed Data (mean of 100 iterations)	0.129	0.098	0.039		
Imputed Data (min)	0.119	0.092	0.034		
Imputed Data (max)	0.140	0.105	0.044		
Panel C: Oregon					
Actual Data	0.097	0.092	0.060	0.081	0.075
Imputed Data (mean of 100 iterations)	0.092	0.087	0.057	0.067	0.068
Imputed Data (min)	0.086	0.082	0.050	0.057	0.059
Imputed Data (max)	0.096	0.093	0.065	0.077	0.076
Student Characteristics		✓	✓	✓	✓
Prior-Year Test Scores			✓	✓	
New-to-School Students Only				✓	
Prior-School Test Scores					✓

Notes: Cells report standard deviations of principal variance components estimated from Equation 8. Actual Data reports the estimate shown in Table A.1. To run our falsification tests, we reassign each school to have a different principal assignment history based on the observed history from another school in the dataset (picked at random). We then estimate our variance decomposition models using these imputed principal assignments. We report the mean across 100 iterations for the estimated principal variance components (in *SD* units). We also report the minimum and maximum estimate across the 100 iterations.

Table A.7: 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	0.296***	0.301***		0.224***		0.031		0.021
	(0.025)	(0.027)		(0.025)		(0.018)		(0.017)
Δ Reading Value-Added		0.079		0.077	0.156***	0.170***		0.122***
		(0.054)		(0.052)	(0.036)	(0.040)		(0.036)
Δ Teacher Experience			0.000	-0.001			-0.001	-0.001
			(0.001)	(0.002)			(0.001)	(0.001)
N	10091	8338	13747	8325	9998	8338	13784	8325

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.8: 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***		0.070***		0.053**		0.049**
	(0.018)	(0.020)		(0.018)		(0.018)		(0.017)
Δ Reading Value-Added		-0.020		-0.013	0.002	-0.023		-0.031
		(0.024)		(0.021)	(0.020)	(0.023)		(0.020)
Δ Teacher Experience			0.000	-0.000			0.004***	0.004***
			(0.001)	(0.001)			(0.001)	(0.001)
N	15436	14961	16614	15085	15330	14967	16613	15085

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.9: 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	0.213***	0.195**		0.155**		0.013		0.028
	(0.047)	(0.063)		(0.054)		(0.062)		(0.053)
Δ Reading Value-Added		0.189**		0.177**	0.213***	0.223***		0.170**
		(0.064)		(0.059)	(0.058)	(0.065)		(0.058)
Δ Teacher Experience			0.001	0.005*			-0.001	0.005*
			(0.001)	(0.002)			(0.001)	(0.002)
N	2038	1410	9706	1411	1730	1408	9704	1411

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 observa-

tions 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.

C Sorting Bias

Equation 6 demonstrates that $\sigma_Y^2(r_x^{\text{SamePrin}} - r_x^{\text{DiffPrin}})$ is a lower-bound estimate of the magnitude of principal effects. This bias arises when schools tend to hire principals of similar quality. A useful expression of the bias is as a proportion of the true *SD* of principal effects, which we call relative bias:

$$\text{Relative Bias} \equiv \frac{\sqrt{\sigma^2_{\delta^F} + \sigma_{\delta^D_x} - \sigma_{\delta^F_x}} - \sqrt{\sigma^2_{\delta^F} + \sigma_{\delta^D_x}}}{\sqrt{\sigma^2_{\delta^F} + \sigma_{\delta^D_x}}} = \sqrt{1 - \rho_\delta(\frac{\sigma^2_{\delta^F}}{\sigma^2_{\delta^F} + \sigma_{\delta^D_x}})} - 1 \qquad (11)$$

Equation 11 shows that, for a given true magnitude of principal effects, bias increases with greater sorting and when principal effectiveness is more stable across years.

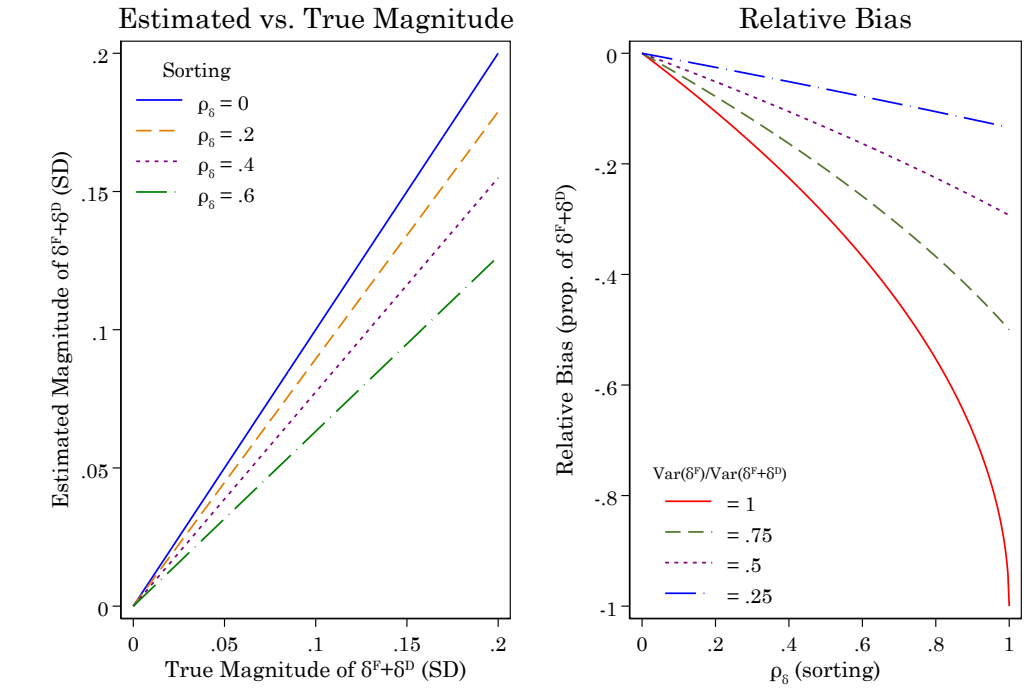


Figure C.1: Sorting Bias

Notes: Left plot shows the estimated magnitude (student test score SD) of principal effects (δ^F) as a function of the true magnitude and the degree of positive sorting bias (ρ_δ), which is the intra-school correlation of δ^F. Here, we assume that principal quality is perfectly stable. The right plot shows the relative bias defined by Equation 11 based on the sorting parameter (ρ_δ) and the stability of principal quality ($\frac{\sigma^2_{\delta^F}}{\sigma^2_{\delta^F}+\sigma_{\delta^D_x}}$). The relative bias shows in proportional terms the downward bias of the lower-bound estimate relative to the true magnitude of principal effects.

Figure C.1 plots the magnitude of sorting bias across a range of scenarios. The left panel shows the true magnitude (in *SD* terms) of principal effects based on the estimated magnitude and the sorting parameter ρ_δ. Here, we assume that principal quality is perfectly stable ($\frac{\sigma^2_{\delta^F}}{\sigma^2_{\delta^F}+\sigma_{\delta^D_x}} = 1$), which is a conservative assumption (relaxing it will decrease the

magnitude of bias). This panel illustrates how the downward bias scales with the true magnitude of principal effects. Put another way, when the estimated magnitude of principal effects is small, the bias (in absolute terms) is small. The right panel plots relative bias (the downward bias expressed as a proportion of the true magnitude of principal effects) based on $\frac{\sigma^2_{\delta F}}{\sigma^2_{\delta F}+\sigma_{\delta^D_x}}$ and ρ_δ. This panel establishes that bias decreases when principal quality is not perfectly stable.

Table C.1: Estimates of ρ_δ Using Alternative Measures of Principal Qualifications and Effectiveness

Measure	N_{schools}	$N_{\text{prin-by-school}}$	Sample Mean	ICC (ρ_δ)
Panel A: Tennessee				
Years of Principal Experience	1829	4027	1.17	.088
Years of Assistant Principal Experience	1946	5011	2.21	.046
Years of Experience (All Roles)	1851	4774	20.9	.044
Has Ph.D./Ed.D./Ed.S. Degree	1850	4781	.415	.106
Average Teacher Rating	1809	3591	-.058	.144
Average Teacher Rating (leave-out)	949	1318	-.100	.054
Average Supervisor Rating	1779	3545	-.102	.285
Average Supervisor Rating (leave-out)	944	1315	-.105	.216
Average Value-Added as Teacher	854	1136	.017	.171
Panel B: New York City				
Years of Principal Experience	1912	5019	1.47	0
Years of Assistant Principal Experience	1588	3068	5.35	.023
Years of Experience (All Roles)	1925	5168	24.4	.09
Average Teacher Rating	1913	2493	-.012	.324
Average Value-Added as Teacher	771	940	.039	0
Panel C: Oregon				
Ever Served as Assistant Principal	1332	3678	.344	.056
Years of Experience (All Roles)	1332	3678	18.9	.062
Has Ph.D./Ed.D. Degree	1332	3678	.045	.083
Does Not Have Master's Degree	1332	3678	.050	.192

Notes: In each row, we estimate the school-level intraclass correlation (ICC) via a random effects variance decomposition model. The ICC is calculated as the school-level variance component divided by the total variance. We also report information about the sample for each measure, including the number of schools, the number of principal-by-school observations, and the sample mean for the measure.

We can bound our estimated magnitude of principal effects under different assumptions about sorting. While the true magnitude of sorting is unknown, we can provide suggestive evidence by estimating ρ_δ for alternative measures of principal quality and observable characteristics, such as experience. Table C.1 presents these estimates. Table C.2 computes bias-corrected estimates of the magnitude of principal effects by applying Equation 11 to our lower-bound estimates from Table 2 and 3. See Section 5.4.1 in the main text for a discussion of these results.

Table C.2: Bias-Corrected Estimates of the Standard Deviation of Principal Effects

$\left(\frac{\sigma^2_{\delta F}}{\sigma^2_{\delta F}+\sigma_{\delta^D_x}}\right)$	ρ_δ = 0.0	0.2	0.4	0.6
Panel A: Math				
1.00	0.034	0.038	0.044	0.054
0.75	0.034	0.037	0.031	0.046
0.50	0.034	0.036	0.038	0.041
Panel B: Reading				
1.00	0.011	0.012	0.014	0.017
0.75	0.011	0.012	0.013	0.015
0.50	0.011	0.012	0.012	0.013

Notes: This table shows bias corrected estimates of the magnitude of principal effects based on applying Equation 11 to our baseline estimates from Tables 2 and 3. To obtain the baseline estimates (represented here in the column where $\rho_\delta = 0$), we compute the simple average of the pooled *SD* of principal effects across each of the three contexts. This average is 0.034 *SD* in math and 0.011 *SD* in reading.

D Examining Stationarity

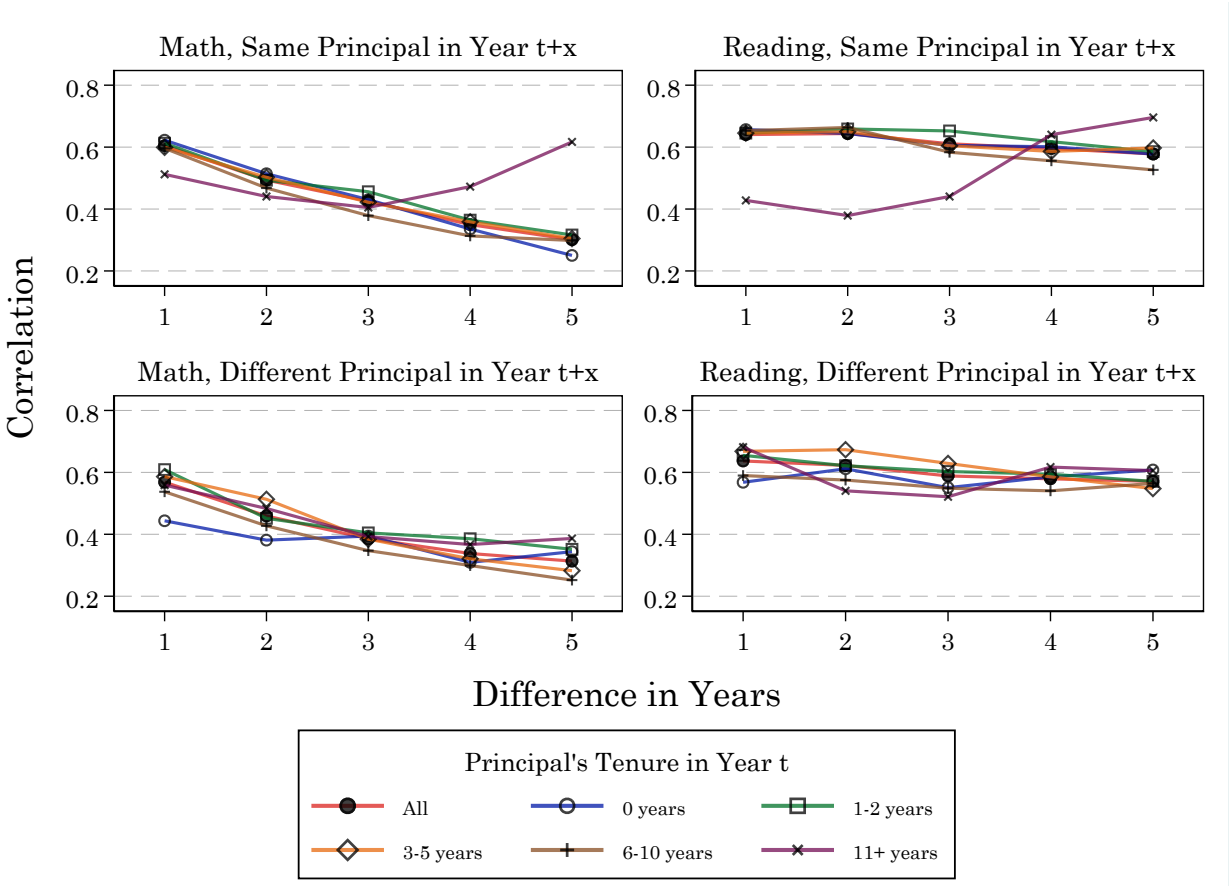


Figure D.1: Autocorrelation Vectors by Current Principal's Tenure (Tennessee)

Notes: Figures report autocorrelation "drift" vectors generated from sample-size-precision-weighted correlations of school-by-year mean residualized test scores generated from Equation 7 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 D.1 shows sample sizes for each of the cells.

Table D.1: Sample Sizes for Figure D.1

		Principal's Tenure in Year t				
	All	0	1–2	3–5	6–10	11+
Panel A: Same Principal						
1 year	11846	2146	3414	3042	2762	482
2 years	8429	1564	2458	2129	2006	272
3 years	6038	1140	1762	1518	1460	158
4 years	4670	900	1377	1152	1073	168
5 years	3181	604	953	766	766	92
Panel B: Different Principal						
1 year	2460	267	688	721	658	126
2 years	4023	509	1154	1124	1075	161
3 years	4928	695	1447	1304	1341	141
4 years	5853	894	1693	1517	1520	229
5 years	5758	934	1612	1515	1504	193

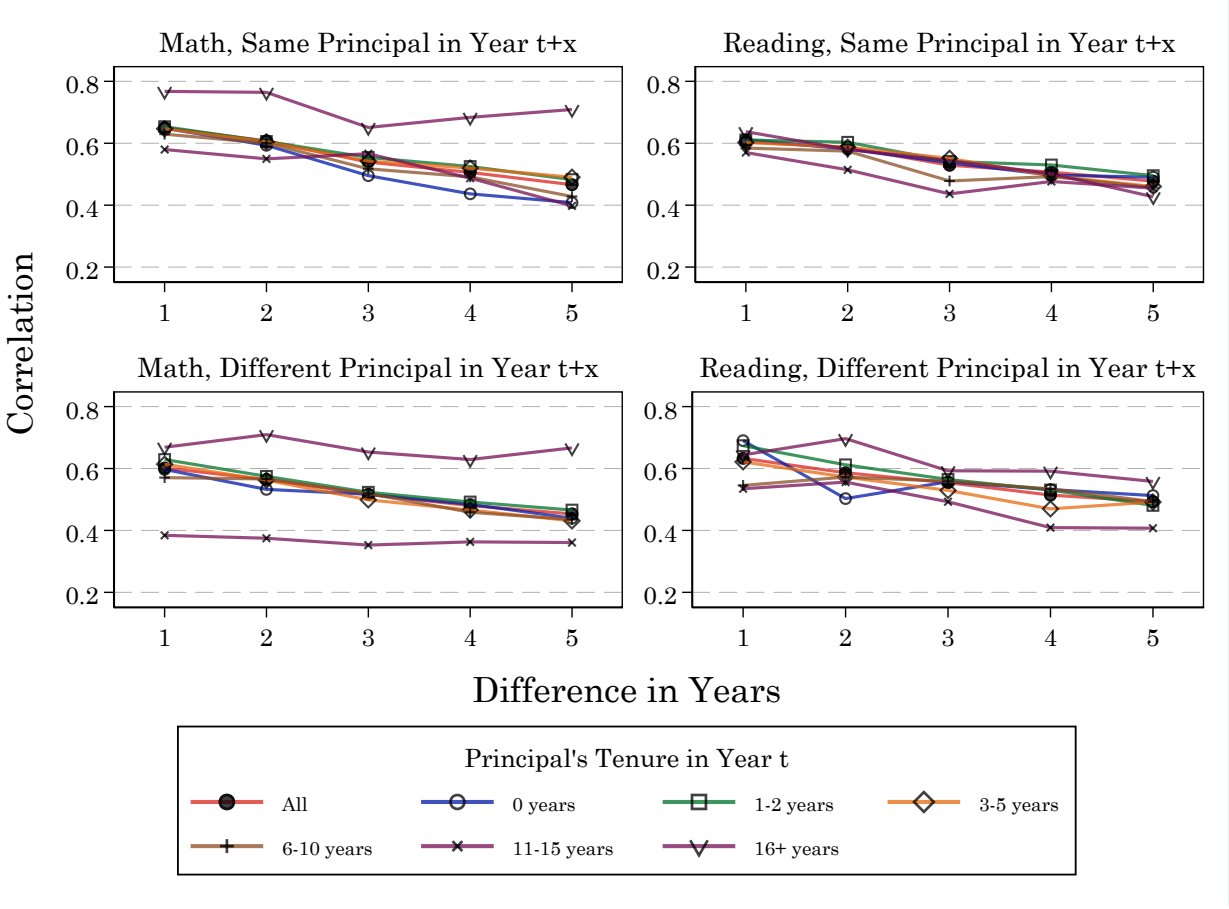


Figure D.2: Autocorrelation Vectors by Current Principal's Tenure (New York City)

Notes: Figures report autocorrelation "drift" vectors generated from sample-size-precision-weighted correlations of school-by-year mean residualized test scores generated from Equation 7 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 D.2 shows sample sizes for each of the cells.

Table D.2: Sample Sizes for Figure D.2

		Principal's Tenure in Year t					
	All	0	1–2	3–5	6–10	11–15	16+
Panel A: Same Principal							
1 year	15145	2328	3974	4009	3552	994	288
2 years	12150	1991	3234	3278	2788	660	199
3 years	9641	1619	2645	2688	2115	432	142
4 years	7629	1319	2173	2215	1538	284	100
5 years	6034	1086	1805	1799	1068	203	73
Panel B: Different Principal							
1 year	2056	273	407	546	532	216	82
2 years	3755	426	848	987	980	372	142
3 years	5013	617	1186	1295	1276	465	174
4 years	5830	764	1403	1530	1439	501	193
5 years	6253	861	1528	1690	1474	500	200

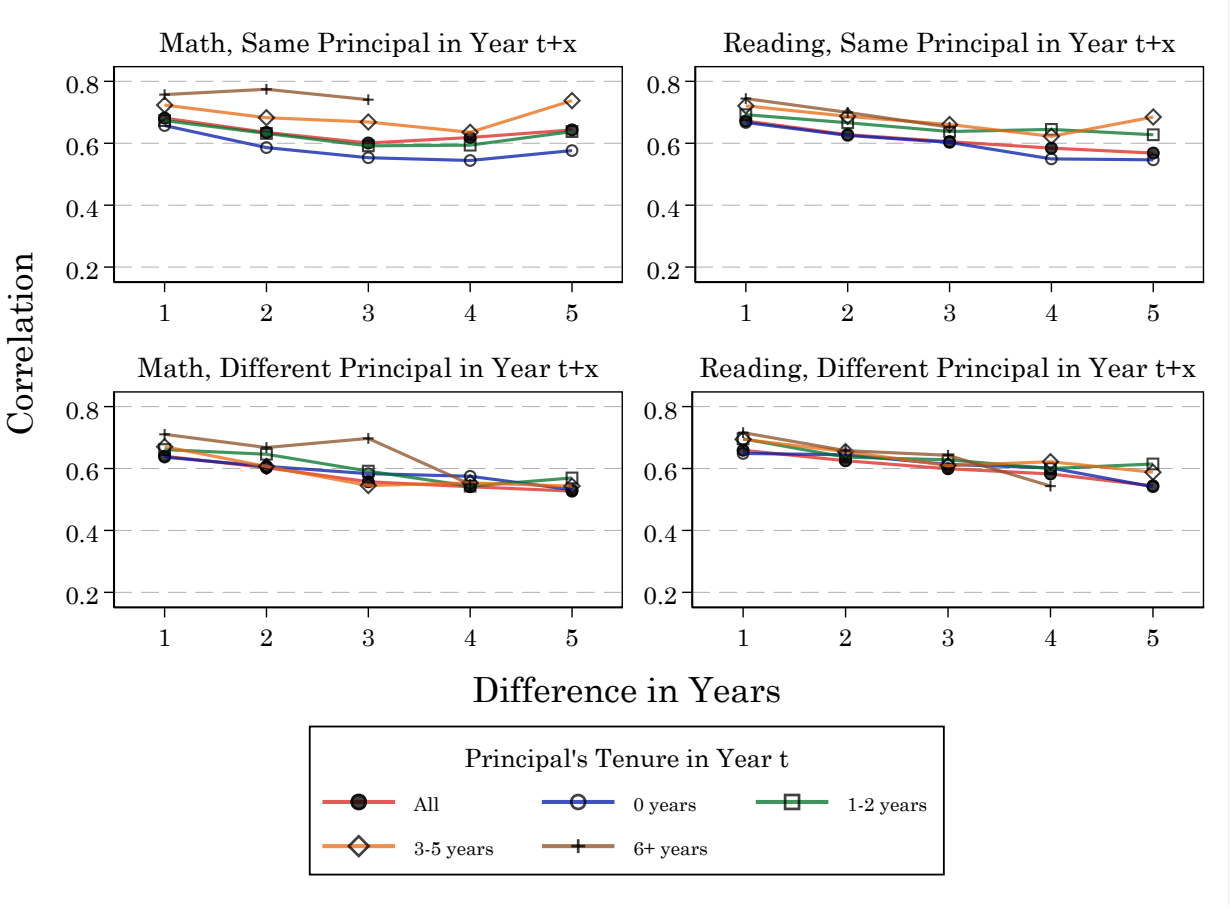


Figure D.3: Autocorrelation Vectors by Current Principal's Tenure (Oregon)

Notes: Figures report autocorrelation "drift" vectors generated from sample-size-precision-weighted correlations of school-by-year mean residualized test scores generated from Equation 7 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 D.3 shows sample sizes for each of the cells.

Table D.3: Sample Sizes for Figure D.3

		Principal's Tenure in Year t			
	All	0	1–2	3–5	6+
Panel A: Same Principal					
1 year	8066	1911	2435	1434	391
2 years	5543	1426	1676	910	208
3 years	3675	1003	1111	549	110
4 years	2374	692	714	324	45
5 years	1470	438	457	182	10
Panel B: Different Principal					
1 year	2004	307	516	419	133
2 years	3374	556	878	666	180
3 years	4180	732	1102	735	164
4 years	4506	825	1180	719	114
5 years	4475	881	1128	635	46

E Balance Checks and Matching Details

Tables E.1, E.2, and E.3 show the mean characteristics of school-by-year cells that comprise the samples to estimate r_x^{same} and r_x^{diff}, respectively, from Tables 2 and 3. See Section 5.4.3 for a description of these results.

To construct samples of r_x^{same} and r_x^{diff} that are balanced on observables, we employ a coarsened exact matching approach (Blackwell et al. 2009). Specifically, we define bins based on the following school-by-year level variables: percentage of FRPL students, percentage of white students, mean test score residuals using students' most recent prior-school score (when estimating the difference in correlations for math [reading], we use the prior-school math [reading] residual). We do not include the test score variable for NYC because we cannot directly observe student enrollment events. We chose these variables because they were those where we observed imbalance in the unmatched comparisons. For each of these four variables, we define five bins. For % FRPL and % white, we construct bins for 0–20, 20–40, 40–60, 60–80, 80–100%. For prior-school test score residuals, we use quintiles of the observed distribution.

Because estimating r_x^{same} and r_x^{diff} requires data from year t and year $t+x$, we include the above measures for both year t and $t+x$. To implement the matching procedure, we define a series of indicators that take a value of 1 if the school's principal in year t is the same as in year $t+x$, and 0 if the principal is different. For each value of x (up to 10 years), we obtain the set of matched "treatment" and "comparison" schools using the strata defined by each combination of the coarsened (binned) variables. Any strata that contain only treatment or comparison schools are effectively dropped (i.e., no common support). The matching algorithm (via the "cem" package in Stata) yields a set of weights (separately for each value of x) that we multiply by our precision weights (the combined sample size of school-by-year cells in year t and $t+x$) to estimate r_x^{same} and r_x^{diff}. Treatment and comparison schools do not differ substantially, on average, in terms of the precision weights, meaning that multiplying the matching weights by the precision weights still yields samples that are observably similar.

Although OR was the only context where we observed substantial differences between r_x^{same} and r_x^{diff} schools, we implement the matching procedure in each context. The following tables show the results. In each context, we show both the balance checks (to illustrate that the matching approach was successful in yielding observably similar sets of schools) and the estimates of r_x^{same} and r_x^{diff} using the matched samples. Tables E.4, E.6, and E.8 show the balance results. For ease of comparison, Tables E.5, E.7, and E.9 show both the baseline (unmatched) and matched results for r_x^{same} and r_x^{diff}, along with the cell sizes. The smaller cell sizes in the matched samples reflect the common support restriction.

Tables E.4, E.6, E.8 demonstrate that our matching approach was successful in reducing observable differences across r_x^{same} and r_x^{diff} samples. In OR, where we observed increasing imbalance with x in the baseline results, Table E.8 shows that r_x^{same} and r_x^{diff} matched samples are now observably similar for all x. Table E.9 shows that the increasing estimated magnitude of principal effects in OR based on $r_x^{\text{same}} - r_x^{\text{diff}}$ does not hold using the matched samples. The magnitude is similar to the baseline results for smaller values of x (where the samples were observably similar even in the baseline results), but is attenuated for larger values of x. For the largest x (i.e., 8 years or more), we actually estimate a negative difference in

correlations (as opposed to a relatively large positive difference in the baseline results), but our precision is quite low. Overall, we interpret these results as suggestive evidence that the larger estimates of the magnitude of principal effects in OR are a function of both sample selection (i.e., an increasingly idiosyncratic set of schools that keep their principal for a long period) and imprecision.

In TN and NYC, the $r_x^{\text{same}} - r_x^{\text{diff}}$ results are very similar between the baseline and matched sample results, which is perhaps unsurprising given that the baseline samples were quite similar on observables.

Table E.1: Mean School Characteristics by Same vs. Different Principal Pairs (Tennessee)

	Prior Math		Prior Read		Attend Rate		White Stu		FRPL Stu		Enroll		Tch Exp		Tch VA	
x	same_x	diff_x	same_x	diff_x	same_x	diff_x	same_x	diff_x	same_x	diff_x	same_x	diff_x	same_x	diff_x	same_x	diff_x
1 year	-0.060	-0.071	-0.040	-0.058	0.947	0.945	0.684	0.646	0.476	0.507	9.178	8.889	12.104	11.860	-0.007	-0.007
2 years	-0.052	-0.064	-0.030	-0.052	0.948	0.946	0.689	0.655	0.483	0.511	9.329	9.084	12.197	11.858	-0.007	-0.009
3 years	-0.049	-0.060	-0.023	-0.045	0.948	0.945	0.694	0.659	0.481	0.509	9.456	9.227	12.310	11.895	-0.007	-0.009
4 years	-0.054	-0.067	-0.024	-0.045	0.949	0.946	0.698	0.661	0.470	0.499	9.433	9.228	12.407	11.953	-0.007	-0.008
5 years	-0.057	-0.066	-0.021	-0.041	0.949	0.946	0.703	0.663	0.463	0.490	9.564	9.394	12.535	11.994	-0.008	-0.009
6 years	-0.050	-0.065	-0.013	-0.036	0.949	0.946	0.705	0.664	0.449	0.480	9.709	9.477	12.636	12.042	-0.008	-0.010
7 years	-0.044	-0.063	-0.005	-0.034	0.950	0.946	0.708	0.663	0.431	0.465	9.986	9.569	12.673	12.043	-0.009	-0.011
8 years	-0.037	-0.057	-0.005	-0.031	0.950	0.945	0.711	0.662	0.406	0.441	10.124	9.763	12.694	12.040	-0.006	-0.012
9 years	-0.064	-0.064	-0.022	-0.033	0.951	0.944	0.717	0.662	0.367	0.411	10.152	10.120	12.700	11.988	-0.003	-0.013
10 years	-0.079	-0.070	-0.033	-0.036	0.952	0.946	0.718	0.662	0.361	0.397	9.659	9.741	12.610	11.925	0.001	-0.011

Notes: Table reports precision-weighted means of school characteristics based on the school-by-year samples that contribute to the estimates of r^{same} and r^{diff} in Table 2. Prior Math and Prior Read are school-by-year mean test score residuals (via residualization Model 1) using each student's most recent prior-year test score in a different school. Attend Rate is the school-by-year mean attendance rate. White Stu and FRPL Stu are the school-by-year proportion of White and free/reduced-price lunch eligible students, respectively. Enroll is the number of students enrolled in the school divided by 100. Tch Exp is the school-by-year mean years of experience of the teaching staff. Tch VA is the school-by-year mean of drift-adjusted, leave-current-school-out value-added. To obtain the mean for a specific cell, we first obtain the precision-weighted mean of the school-by-year means for the current year and future year (e.g., diff = 3 years includes year t and year $t+3$) among pairs of school-by-year cells that are included in the calculations of r^{same} and r^{diff}, respectively, where the weights are those used to calculate r^{same} and r^{diff}. We then take the simple average of those two means (i.e., $(\bar{x}_t + \bar{x}_{t+3})/2$).

Table E.2: Mean School Characteristics by Same vs. Different Principal Pairs (New York City)

	Attend Rate		White Stu		FRPL Stu		Enroll		Tch Exp		Tch VA	
x	same_x	diff_x	same_x	diff_x	same_x	diff_x	same_x	diff_x	same_x	diff_x	same_x	diff_x
1 year	92.200	91.398	15.314	13.817	0.822	0.809	9.192	9.404	10.794	10.650	-0.016	-0.020
2 years	92.445	91.618	15.551	14.004	0.826	0.809	9.157	9.273	10.799	10.570	-0.015	-0.019
3 years	92.634	91.846	15.773	14.422	0.829	0.809	9.108	9.231	10.860	10.582	-0.013	-0.018
4 years	92.852	92.074	15.967	14.787	0.833	0.810	9.060	9.162	10.962	10.659	-0.013	-0.017
5 years	93.026	92.254	15.946	15.230	0.844	0.817	9.042	9.071	11.099	10.754	-0.012	-0.016
6 years	93.175	92.423	16.029	15.531	0.859	0.837	9.044	8.998	11.269	10.915	-0.012	-0.015
7 years	93.312	92.583	16.298	15.697	0.860	0.840	9.092	8.957	11.451	11.110	-0.012	-0.015
8 years	93.443	92.789	16.423	15.946	0.863	0.848	9.172	8.961	11.616	11.320	-0.011	-0.014
9 years	93.587	93.013	16.799	16.073	0.869	0.860	9.258	9.006	11.746	11.535	-0.010	-0.014
10 years	93.728	93.202	17.202	16.200	0.874	0.866	9.333	9.087	11.870	11.748	-0.009	-0.014

Notes: Table reports precision-weighted means of school characteristics based on the school-by-year samples that contribute to the estimates of r^{same} and r^{diff} in Table 2. Prior Math and Prior Read are school-by-year mean test score residuals (via residualization Model 1) using each student's most recent prior-year test score in a different school. Attend Rate is the school-by-year mean attendance rate. White Stu and FRPL Stu are the school-by-year proportion of White and free/reduced-price lunch eligible students, respectively. Enroll is the number of students enrolled in the school divided by 100. Tch Exp is the school-by-year mean years of experience of the teaching staff. Tch VA is the school-by-year mean of drift-adjusted, leave-current-school-out value-added. To obtain the mean for a specific cell, we first obtain the precision-weighted mean of the school-by-year means for the current year and future year (e.g., diff = 3 years includes year t and year $t+3$) among pairs of school-by-year cells that are included in the calculations of r^{same} and r^{diff}, respectively, where the weights are those used to calculate r^{same} and r^{diff}. We then take the simple average of those two means (i.e., $(\bar{x}_t + \bar{x}_{t+3})/2$).

Table E.3: Mean School Characteristics by Same vs. Different Principal Pairs (Oregon)

	Prior Math		Prior Read		Attend Rate		White Stu		FRPL Stu		Enroll		Tch Exp		Tch VA	
x	same_x	diff_x	same_x	diff_x	same_x	diff_x	same_x	diff_x	same_x	diff_x	same_x	diff_x	same_x	diff_x	same_x	diff_x
1 year	-0.012	-0.019	-0.006	-0.007	0.944	0.943	0.645	0.641	0.511	0.519	7.915	8.087	12.446	12.238	-0.013	-0.014
2 years	-0.010	-0.020	-0.005	-0.009	0.945	0.944	0.642	0.632	0.520	0.532	7.890	8.163	12.539	12.343	-0.013	-0.017
3 years	-0.004	-0.019	-0.003	-0.006	0.945	0.944	0.640	0.625	0.524	0.542	7.865	8.168	12.611	12.434	-0.009	-0.019
4 years	0.010	-0.024	0.006	-0.011	0.946	0.944	0.639	0.619	0.518	0.549	7.730	8.180	12.644	12.429	-0.011	-0.014
5 years	0.022	-0.025	0.012	-0.012	0.947	0.945	0.636	0.616	0.516	0.556	7.834	7.989	12.615	12.404	-0.011	-0.012
6 years	0.048	-0.022	0.024	-0.009	0.947	0.945	0.632	0.614	0.511	0.558	7.924	7.862	12.591	12.305		
7 years	0.060	-0.019	0.023	-0.007	0.947	0.945	0.632	0.610	0.503	0.563	7.928	7.698	12.488	12.141		
8 years	0.084	-0.017	0.031	-0.007	0.948	0.945	0.628	0.607	0.491	0.570	7.568	7.405	12.417	12.024		
9 years	0.093	-0.017	0.041	-0.008	0.948	0.944	0.626	0.604	0.458	0.574	7.449	7.367	12.396	12.001		
10 years	0.089	-0.016	0.050	-0.006	0.947	0.943	0.623	0.601	0.429	0.573	7.899	7.337	12.425	11.989		

Notes: Table reports precision-weighted means of school characteristics based on the school-by-year samples that contribute to the estimates of r^{same} and r^{diff} in Table 2. Prior Math and Prior Read are school-by-year mean test score residuals (via residualization Model 1) using each student's most recent prior-year test score in a different school. Attend Rate is the school-by-year mean attendance rate. White Stu and FRPL Stu are the school-by-year proportion of White and free/reduced-price lunch eligible students, respectively. Enroll is the number of students enrolled in the school divided by 100. Tch Exp is the school-by-year mean years of experience of the teaching staff. Tch VA is the school-by-year mean of drift-adjusted, leave-current-school-out value-added. To obtain the mean for a specific cell, we first obtain the precision-weighted mean of the school-by-year means for the current year and future year (e.g., diff = 3 years includes year t and year $t+3$) among pairs of school-by-year cells that are included in the calculations of r^{same} and r^{diff}, respectively, where the weights are those used to calculate r^{same} and r^{diff}. We then take the simple average of those two means (i.e., $(\bar{x}_t + \bar{x}_{t+3})/2$).

Table E.4: Mean School Characteristics by Same vs. Different Principal Pairs (Tennessee, Matched Sample)

	Prior Math		Prior Read		Attend Rate		White Stu		FRPL Stu		Enroll		Tch Exp		Tch VA	
x	same_x	diff_x	same_x	diff_x	same_x	diff_x	same_x	diff_x	same_x	diff_x	same_x	diff_x	same_x	diff_x	same_x	diff_x
Panel A: Matched Sample for Math																
1 year	-0.071	-0.066	-0.051	-0.050	0.948	0.946	0.690	0.694	0.486	0.484	9.175	8.989	12.108	12.060	-0.007	-0.008
2 years	-0.064	-0.064	-0.042	-0.048	0.948	0.947	0.695	0.700	0.497	0.500	9.268	9.066	12.131	11.944	-0.007	-0.009
3 years	-0.065	-0.063	-0.039	-0.040	0.948	0.946	0.698	0.701	0.492	0.491	9.348	9.204	12.170	11.860	-0.006	-0.009
4 years	-0.066	-0.067	-0.032	-0.034	0.949	0.948	0.697	0.704	0.473	0.470	9.253	9.182	12.128	11.775	-0.007	-0.008
5 years	-0.066	-0.070	-0.027	-0.032	0.949	0.947	0.699	0.701	0.461	0.463	9.280	9.182	12.173	11.716	-0.007	-0.009
6 years	-0.063	-0.070	-0.021	-0.025	0.950	0.948	0.700	0.701	0.443	0.445	9.407	9.216	12.182	11.739	-0.008	-0.008
7 years	-0.059	-0.055	-0.015	-0.010	0.950	0.949	0.705	0.702	0.415	0.415	9.644	9.429	12.167	11.713	-0.008	-0.008
8 years	-0.063	-0.076	-0.024	-0.031	0.950	0.948	0.715	0.710	0.381	0.390	9.699	9.178	12.227	11.758	-0.007	-0.011
9 years	-0.090	-0.080	-0.041	-0.036	0.952	0.949	0.722	0.705	0.374	0.379	8.874	9.100	12.456	11.819	-0.003	-0.010
10 years	-0.109	-0.086	-0.059	-0.033	0.954	0.950	0.733	0.720	0.315	0.315	8.692	9.025	12.582	11.918	0.001	-0.012
Panel B: Matched Sample for Reading																
1 year	-0.053	-0.062	-0.036	-0.047	0.947	0.945	0.690	0.696	0.480	0.480	9.644	9.467	12.124	12.061	-0.008	-0.008
2 years	-0.046	-0.055	-0.027	-0.041	0.947	0.945	0.694	0.700	0.491	0.494	9.751	9.599	12.136	11.969	-0.008	-0.011
3 years	-0.045	-0.055	-0.022	-0.035	0.947	0.945	0.697	0.704	0.485	0.487	9.838	9.665	12.171	11.884	-0.008	-0.010
4 years	-0.048	-0.061	-0.019	-0.030	0.948	0.947	0.700	0.711	0.465	0.468	9.714	9.451	12.144	11.812	-0.009	-0.010
5 years	-0.057	-0.066	-0.023	-0.035	0.948	0.946	0.704	0.709	0.458	0.463	9.760	9.590	12.135	11.739	-0.009	-0.011
6 years	-0.045	-0.065	-0.007	-0.029	0.949	0.947	0.703	0.697	0.437	0.448	9.880	9.603	12.134	11.713	-0.009	-0.010
7 years	-0.042	-0.063	-0.001	-0.027	0.950	0.947	0.706	0.697	0.412	0.426	10.069	9.694	12.128	11.683	-0.008	-0.010
8 years	-0.042	-0.066	-0.007	-0.029	0.950	0.947	0.712	0.709	0.376	0.386	10.129	9.606	12.200	11.809	-0.007	-0.013
9 years	-0.074	-0.085	-0.026	-0.042	0.951	0.948	0.724	0.712	0.367	0.377	9.277	9.273	12.449	11.843	-0.003	-0.010
10 years	-0.087	-0.085	-0.040	-0.037	0.953	0.950	0.729	0.725	0.309	0.316	9.015	8.924	12.513	11.943	0.001	-0.011

Notes: Table reports precision-weighted means of school characteristics based on the school-by-year samples that contribute to the estimates of r^{same} and r^{diff} in Table 2. Prior Math and Prior Read are school-by-year mean test score residuals (via residualization Model 1) using each student's most recent prior-year test score in a different school. Attend Rate is the school-by-year mean attendance rate. White Stu and FRPL Stu are the school-by-year proportion of White and free/reduced-price lunch eligible students, respectively. Enroll is the number of students enrolled in the school divided by 100. Tch Exp is the school-by-year mean years of experience of the teaching staff. Tch VA is the school-by-year mean of drift-adjusted, leave-current-school-out value-added. To obtain the mean for a specific cell, we first obtain the precision-weighted mean of the school-by-year means for the current year and future year (e.g., diff = 3 years includes year t and year $t+3$) among pairs of school-by-year cells that are included in the calculations of r^{same} and r^{diff}, respectively, where the weights are those used to calculate r^{same} and r^{diff}. We then take the simple average of those two means (i.e., $(\bar{x}_t + \bar{x}_{t+3})/2$).

Table E.5: Differences in Autocorrelations Before and After Matching (Tennessee)

	Baseline						Matched Sample					
x	N_x^{same}	N_x^{diff}	r_x^{same}	r_x^{diff}	diff_x	SD_x	N_x^{same}	N_x^{diff}	r_x^{same}	r_x^{diff}	diff_x	SD_x
Panel A: Math												
1 year	11846	2460	0.605	0.569	0.036	0.035	10754	2223	0.609	0.578	0.031	0.032
2 years	8429	4023	0.494	0.459	0.034	0.034	7761	3691	0.487	0.473	0.015	0.022
3 years	6038	4928	0.425	0.386	0.040	0.036	5584	4506	0.424	0.397	0.027	0.030
4 years	4670	5853	0.350	0.338	0.012	0.020	4308	5319	0.354	0.344	0.009	0.018
5 years	3181	5758	0.301	0.313	-0.012	0.000	2929	5130	0.306	0.325	-0.019	0.000
6 years	2125	5254	0.272	0.244	0.028	0.030	1942	4503	0.272	0.256	0.015	0.022
7 years	1312	4545	0.193	0.189	0.004	0.012	1183	3643	0.209	0.212	-0.003	0.000
8 years	786	3589	0.229	0.157	0.072	0.049	700	2887	0.232	0.192	0.041	0.037
9 years	599	3416	0.230	0.185	0.045	0.039	522	2666	0.253	0.207	0.046	0.039
10 years	340	2426	0.331	0.143	0.188	0.079	288	1783	0.332	0.164	0.168	0.075
Pooled *SD* Estimate						0.032						0.026
Panel B: Reading												
1 year	11880	2468	0.640	0.637	0.004	0.008	11114	2254	0.633	0.633	0.000	0.001
2 years	8457	4038	0.644	0.622	0.021	0.019	7990	3739	0.640	0.634	0.006	0.010
3 years	6056	4947	0.611	0.589	0.022	0.020	5718	4601	0.605	0.596	0.009	0.013
4 years	4685	5875	0.595	0.579	0.016	0.017	4324	5310	0.596	0.585	0.011	0.014
5 years	3189	5775	0.578	0.571	0.007	0.011	2916	5180	0.564	0.577	-0.014	0.000
6 years	2131	5272	0.551	0.550	0.001	0.003	1970	4660	0.551	0.554	-0.004	0.000
7 years	1313	4565	0.542	0.521	0.021	0.019	1220	3910	0.537	0.533	0.004	0.008
8 years	785	3599	0.494	0.481	0.013	0.015	707	2907	0.473	0.496	-0.023	0.000
9 years	598	3420	0.533	0.508	0.026	0.021	525	2683	0.532	0.505	0.027	0.022
10 years	339	2426	0.547	0.444	0.103	0.043	288	1678	0.550	0.487	0.063	0.033
Pooled *SD* Estimate						0.016						0.008

Notes: Table reports original autocorrelation results from Tables 2 and 3 alongside same analysis performed on the matched sample. Samples differ in baseline and matched sample due to lack of common support in particular strata. See Section E for details on matching process and table notes in Table 2 for additional information.

Table E.6: Mean School Characteristics by Same vs. Different Principal Pairs (New York City, Matched Sample)

	Attend Rate		White Stu		FRPL Stu		Enroll		Tch Exp		Tch VA	
x	same$_x$	diff$_x$	same$_x$	diff$_x$	same$_x$	diff$_x$	same$_x$	diff$_x$	same$_x$	diff$_x$	same$_x$	diff$_x$
Panel A: Matched Sample for Math												
1 year	0.922	0.915	0.143	0.133	0.831	0.828	9.072	9.234	10.702	10.549	-0.015	-0.020
2 years	0.924	0.917	0.148	0.139	0.832	0.826	8.993	9.062	10.781	10.603	-0.014	-0.020
3 years	0.927	0.919	0.150	0.139	0.836	0.832	8.886	8.935	10.920	10.657	-0.013	-0.019
4 years	0.928	0.921	0.151	0.142	0.845	0.838	8.852	8.830	11.071	10.763	-0.013	-0.017
5 years	0.930	0.922	0.151	0.142	0.863	0.856	8.822	8.747	11.267	10.920	-0.013	-0.016
6 years	0.931	0.923	0.153	0.145	0.865	0.860	8.830	8.652	11.484	11.120	-0.012	-0.015
7 years	0.932	0.925	0.155	0.146	0.869	0.867	8.859	8.624	11.671	11.358	-0.012	-0.015
8 years	0.934	0.928	0.154	0.145	0.877	0.877	8.878	8.678	11.829	11.586	-0.011	-0.013
9 years	0.935	0.930	0.158	0.146	0.881	0.880	8.960	8.694	11.984	11.813	-0.010	-0.014
10 years	0.936	0.931	0.161	0.146	0.889	0.887	9.002	8.793	12.100	12.001	-0.009	-0.013
Panel B: Matched Sample for Reading												
1 year	0.923	0.915	0.145	0.136	0.831	0.827	9.012	9.130	10.715	10.566	-0.015	-0.020
2 years	0.925	0.917	0.150	0.142	0.832	0.825	8.936	8.975	10.795	10.628	-0.014	-0.020
3 years	0.927	0.919	0.152	0.142	0.835	0.831	8.857	8.875	10.937	10.683	-0.013	-0.018
4 years	0.929	0.921	0.153	0.145	0.844	0.837	8.833	8.784	11.084	10.786	-0.013	-0.017
5 years	0.930	0.922	0.154	0.144	0.861	0.855	8.810	8.712	11.276	10.940	-0.013	-0.016
6 years	0.931	0.924	0.155	0.147	0.863	0.859	8.823	8.625	11.486	11.132	-0.012	-0.015
7 years	0.933	0.926	0.157	0.149	0.868	0.866	8.851	8.608	11.667	11.362	-0.012	-0.015
8 years	0.934	0.928	0.157	0.147	0.875	0.874	8.871	8.666	11.822	11.583	-0.012	-0.013
9 years	0.935	0.930	0.161	0.149	0.880	0.878	8.941	8.682	11.977	11.804	-0.010	-0.013
10 years	0.936	0.932	0.165	0.150	0.887	0.884	8.971	8.795	12.097	11.993	-0.009	-0.013

Notes: Table reports precision-weighted means of school characteristics based on the school-by-year samples that contribute to the estimates of r^{same} and r^{diff} in Table 2. Prior Math and Prior Read are school-by-year mean test score residuals (via residualization Model 1) using each student's most recent prior-year test score in a different school. Attend Rate is the school-by-year mean attendance rate. White Stu and FRPL Stu are the school-by-year proportion of White and free/reduced-price lunch eligible students, respectively. Enroll is the number of students enrolled in the school divided by 100. Tch Exp is the school-by-year mean years of experience of the teaching staff. Tch VA is the school-by-year mean of drift-adjusted, leave-current-school-out value-added. To obtain the mean for a specific cell, we first obtain the precision-weighted mean of the school-by-year means for the current year and future year (e.g., diff = 3 years includes year t and year $t+3$) among pairs of school-by-year cells that are included in the calculations of r^{same} and r^{diff}, respectively, where the weights are those used to calculate r^{same} and r^{diff}. We then take the simple average of those two means (i.e., $(\bar{x}_t + \bar{x}_{t+3})/2$).

Table E.7: Differences in Autocorrelations Before and After Matching (New York City)

	Baseline						Matched Sample					
x	N_x^{same}	N_x^{diff}	r_x^{same}	r_x^{diff}	diff_x	SD_x	N_x^{same}	N_x^{diff}	r_x^{same}	r_x^{diff}	diff_x	SD_x
Panel A: Math												
1 year	14848	2015	0.651	0.605	0.046	0.030	14762	2046	0.650	0.597	0.053	0.032
2 years	11900	3693	0.609	0.567	0.042	0.029	11965	3729	0.607	0.567	0.040	0.028
3 years	9437	4924	0.543	0.519	0.024	0.022	9536	4974	0.538	0.518	0.019	0.019
4 years	7462	5728	0.509	0.482	0.027	0.023	7535	5753	0.505	0.479	0.026	0.022
5 years	5897	6136	0.471	0.456	0.014	0.017	5957	6147	0.464	0.449	0.014	0.017
6 years	4629	6278	0.436	0.430	0.006	0.011	4682	6257	0.428	0.426	0.002	0.007
7 years	3597	6201	0.410	0.410	-0.001	0.000	3647	6155	0.409	0.401	0.008	0.012
8 years	2745	5974	0.371	0.384	-0.013	0.000	2776	5901	0.374	0.379	-0.005	0.000
9 years	2019	5656	0.346	0.358	-0.011	0.000	2052	5551	0.350	0.361	-0.011	0.000
10 years	1428	5233	0.333	0.339	-0.006	0.000	1452	5104	0.333	0.336	-0.002	0.000
Pooled *SD* Estimate						0.022						0.023
Panel B: Reading												
1 year	14832	2017	0.608	0.638	-0.030	0.000	14797	2059	0.603	0.636	-0.033	0.000
2 years	11883	3693	0.591	0.599	-0.008	0.000	11963	3734	0.586	0.598	-0.012	0.000
3 years	9422	4926	0.535	0.558	-0.024	0.000	9535	4982	0.528	0.554	-0.027	0.000
4 years	7450	5728	0.513	0.520	-0.007	0.000	7535	5760	0.505	0.511	-0.006	0.000
5 years	5888	6132	0.482	0.505	-0.023	0.000	5954	6146	0.476	0.491	-0.014	0.000
6 years	4623	6272	0.460	0.480	-0.020	0.000	4678	6261	0.453	0.476	-0.023	0.000
7 years	3591	6195	0.444	0.467	-0.023	0.000	3644	6157	0.442	0.455	-0.013	0.000
8 years	2741	5966	0.440	0.471	-0.031	0.000	2774	5904	0.434	0.465	-0.031	0.000
9 years	2016	5650	0.417	0.458	-0.041	0.000	2052	5552	0.412	0.460	-0.048	0.000
10 years	1426	5226	0.413	0.453	-0.039	0.000	1453	5104	0.404	0.446	-0.042	0.000
Pooled *SD* Estimate						0.000						0.000

Notes: Table reports original autocorrelation results from Tables 2 and 3 alongside same analysis performed on the matched sample. Samples differ in baseline and matched sample due to lack of common support in particular strata. See Section E for details on matching process and table notes in Table 2 for additional information.

Table E.8: Mean School Characteristics by Same vs. Different Principal Pairs (Oregon, Matched Sample)

	Prior Math		Prior Read		Attend Rate		White Stu		FRPL Stu		Enroll		Tch Exp		Tch VA	
x	same_x	diff_x	same_x	diff_x	same_x	diff_x	same_x	diff_x	same_x	diff_x	same_x	diff_x	same_x	diff_x	same_x	diff_x
Panel A: Matched Sample for Math																
1 year	-0.012	-0.013	-0.004	-0.003	0.945	0.943	0.643	0.646	0.516	0.518	8.088	8.306	12.513	12.451	-0.015	-0.014
2 years	-0.014	-0.015	-0.008	-0.006	0.945	0.944	0.638	0.638	0.528	0.530	8.013	8.363	12.602	12.509	-0.014	-0.018
3 years	-0.008	-0.009	-0.005	0.002	0.945	0.944	0.634	0.635	0.528	0.530	7.997	8.389	12.595	12.479	-0.010	-0.019
4 years	-0.005	-0.009	-0.004	0.003	0.946	0.944	0.636	0.635	0.524	0.530	7.915	8.374	12.618	12.468	-0.014	-0.014
5 years	-0.005	-0.011	0.000	0.002	0.945	0.944	0.634	0.637	0.526	0.528	7.978	8.259	12.490	12.411	-0.017	-0.015
6 years	0.007	-0.006	0.000	0.002	0.945	0.944	0.633	0.636	0.531	0.528	8.080	8.373	12.464	12.194		
7 years	0.005	0.012	-0.014	0.009	0.945	0.944	0.633	0.641	0.529	0.518	7.952	7.955	12.299	12.200		
8 years	0.008	0.019	-0.026	0.010	0.945	0.943	0.633	0.640	0.515	0.518	7.797	7.632	12.287	12.193		
9 years	0.020	0.047	-0.013	0.043	0.946	0.944	0.638	0.637	0.474	0.491	7.365	7.372	12.446	12.142		
10 years	0.035	0.021	0.017	0.002	0.944	0.942	0.618	0.618	0.457	0.489	7.866	7.926	12.406	12.239		
Panel B: Matched Sample for Reading																
1 year	-0.018	-0.021	-0.008	-0.006	0.945	0.943	0.643	0.644	0.523	0.526	7.897	8.262	12.506	12.388	-0.014	-0.012
2 years	-0.020	-0.015	-0.010	-0.002	0.945	0.944	0.637	0.638	0.534	0.534	7.848	8.181	12.579	12.512	-0.014	-0.017
3 years	-0.023	-0.016	-0.015	-0.002	0.945	0.944	0.635	0.636	0.536	0.536	7.841	8.358	12.581	12.497	-0.010	-0.020
4 years	-0.011	-0.012	-0.006	0.003	0.946	0.944	0.637	0.635	0.529	0.534	7.784	8.262	12.592	12.436	-0.013	-0.015
5 years	-0.003	-0.007	0.000	0.009	0.946	0.945	0.634	0.637	0.528	0.529	7.816	8.020	12.497	12.398	-0.015	-0.015
6 years	0.010	-0.014	0.001	-0.000	0.945	0.944	0.635	0.635	0.532	0.534	7.892	8.219	12.416	12.251		
7 years	0.005	-0.013	-0.018	-0.004	0.945	0.944	0.628	0.635	0.530	0.525	7.783	7.679	12.266	12.139		
8 years	0.015	-0.004	-0.016	-0.011	0.945	0.944	0.632	0.634	0.510	0.524	7.553	7.340	12.284	12.142		
9 years	0.040	0.038	0.011	0.029	0.946	0.944	0.643	0.637	0.462	0.487	7.168	7.338	12.413	12.363		
10 years	0.039	0.034	0.039	0.031	0.945	0.943	0.623	0.626	0.443	0.468	7.610	7.468	12.344	12.452		

Notes: Table reports precision-weighted means of school characteristics based on the school-by-year samples that contribute to the estimates of r^{same} and r^{diff} in Table 2. Prior Math and Prior Read are school-by-year mean test score residuals (via residualization Model 1) using each student's most recent prior-year test score in a different school. Attend Rate is the school-by-year mean attendance rate. White Stu and FRPL Stu are the school-by-year proportion of White and free/reduced-price lunch eligible students, respectively. Enroll is the number of students enrolled in the school divided by 100. Tch Exp is the school-by-year mean years of experience of the teaching staff. Tch VA is the school-by-year mean of drift-adjusted, leave-current-school-out value-added. To obtain the mean for a specific cell, we first obtain the precision-weighted mean of the school-by-year means for the current year and future year (e.g., diff = 3 years includes year t and year $t+3$) among pairs of school-by-year cells that are included in the calculations of r^{same} and r^{diff}, respectively, where the weights are those used to calculate r^{same} and r^{diff}. We then take the simple average of those two means (i.e., $(\bar{x}_t + \bar{x}_{t+3})/2$).

Table E.9: Differences in Autocorrelations Before and After Matching (Oregon)

	Baseline						Matched Sample					
x	N_x^{same}	N_x^{diff}	r_x^{same}	r_x^{diff}	diff_x	SD_x	N_x^{same}	N_x^{diff}	r_x^{same}	r_x^{diff}	diff_x	SD_x
Panel A: Math												
1 year	8066	2004	0.681	0.640	0.041	0.041	7066	1828	0.673	0.637	0.036	0.038
2 years	5543	3374	0.635	0.603	0.032	0.036	4918	3027	0.618	0.605	0.013	0.023
3 years	3675	4180	0.601	0.557	0.044	0.042	3293	3674	0.575	0.557	0.018	0.027
4 years	2374	4506	0.619	0.541	0.077	0.056	2117	3717	0.586	0.581	0.005	0.014
5 years	1470	4475	0.643	0.527	0.116	0.068	1324	3483	0.578	0.555	0.023	0.030
6 years	894	4162	0.647	0.522	0.125	0.071	784	2923	0.571	0.559	0.012	0.022
7 years	534	3678	0.653	0.482	0.170	0.082	455	2239	0.568	0.577	-0.009	0.000
8 years	303	3040	0.675	0.518	0.157	0.079	252	1604	0.571	0.609	-0.038	0.000
9 years	171	2320	0.676	0.514	0.162	0.080	137	872	0.538	0.638	-0.100	0.000
10 years	86	1559	0.573	0.518	0.055	0.047	65	474	0.412	0.645	-0.233	0.000
Pooled *SD* Estimate						0.048						0.027
Panel B: Reading												
1 year	8065	2006	0.672	0.659	0.013	0.023	7239	1836	0.668	0.662	0.006	0.015
2 years	5544	3375	0.628	0.625	0.003	0.010	5030	3045	0.623	0.614	0.008	0.018
3 years	3676	4179	0.605	0.599	0.006	0.015	3327	3653	0.593	0.585	0.009	0.018
4 years	2374	4503	0.584	0.583	0.002	0.008	2123	3752	0.572	0.570	0.002	0.010
5 years	1469	4473	0.568	0.544	0.024	0.031	1340	3632	0.558	0.559	-0.002	0.000
6 years	893	4161	0.565	0.531	0.034	0.037	781	2981	0.561	0.542	0.019	0.027
7 years	535	3682	0.567	0.516	0.051	0.045	452	2239	0.552	0.567	-0.015	0.000
8 years	303	3041	0.567	0.514	0.053	0.046	256	1548	0.563	0.537	0.026	0.032
9 years	171	2320	0.505	0.503	0.002	0.010	136	864	0.439	0.562	-0.123	0.000
10 years	86	1560	0.486	0.490	-0.004	0.000	67	496	0.424	0.573	-0.148	0.000
Pooled *SD* Estimate						0.018						0.013

Notes: Table reports original autocorrelation results from Tables 2 and 3 alongside same analysis performed on the matched sample. Samples differ in baseline and matched sample due to lack of common support in particular strata. See Section E for details on matching process and table notes in Table 2 for additional information.

F Analyzing School-Switching Principals

Our primary decomposition analyses examine differences between principals who worked in the same school but do not examine the stability of school performance across schools led by the same principal. The primary analyses suggest that most of the between-principal variation in school performance likely reflects factors other than principal performance. One way to further check this result is to 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 in Equation 2, this correlation is:

$$r_x^{\text{DiffSch}} = \frac{\sigma^2_{\delta^F} + \sigma_{\delta^D_x} + \sigma_{\mu^F_x}}{\sigma^2_Y}$$
$$\text{where } \sigma_{\mu^F_x} = \text{cov}(\mu^F_{j(p,t)}, \mu^F_{k(p,t+x)}) < \sigma^2_{\mu^F} \tag{12}$$

Given our baseline results, which suggest that $\sigma^2_{\delta^F} + \sigma_{\delta^D_x}$ is small, 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 in terms of fixed factors μ^F that affect student test score performance, $\sigma_{\mu^F_x}$ will be positive and r_x^{DiffSch} will increase.[33]

Figure F.1 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 via $\sigma_{\mu^F_x}$. To show this, we can estimate r_x^{DiffSch} for our teacher and student composition measures. As demonstrated by Appendix Figures F.2 and F.3, 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. Notably, the within-principal, between-school correlation is roughly 0.2 in all contexts and subjects, which is similar to the suggested magnitude of principal sorting from Table C.1.

33. 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).

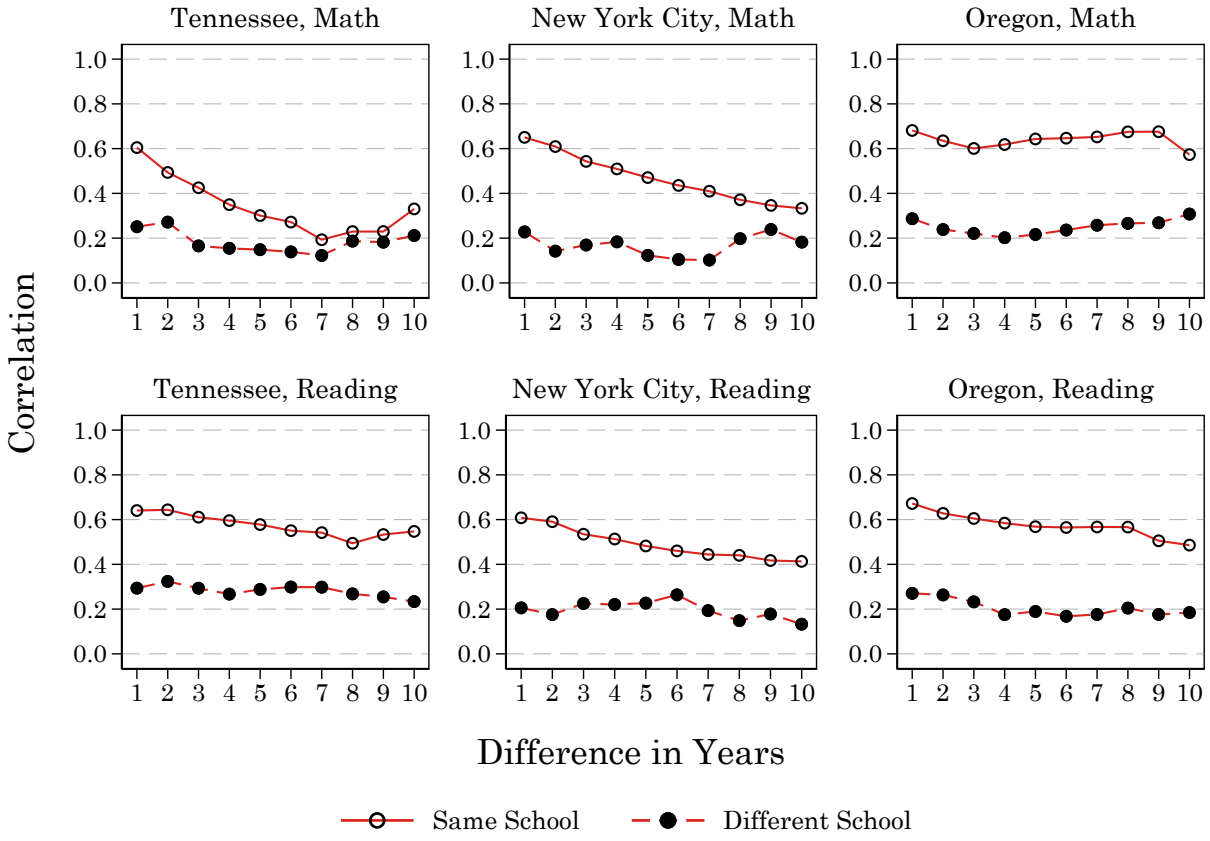


Figure F.1: Autocorrelations Within and Between Schools

Notes: Figures report autocorrelation "drift" vectors generated from sample-size-precision-weighted correlations of principal-by-year mean residualized test score generated from Equation 7. 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).

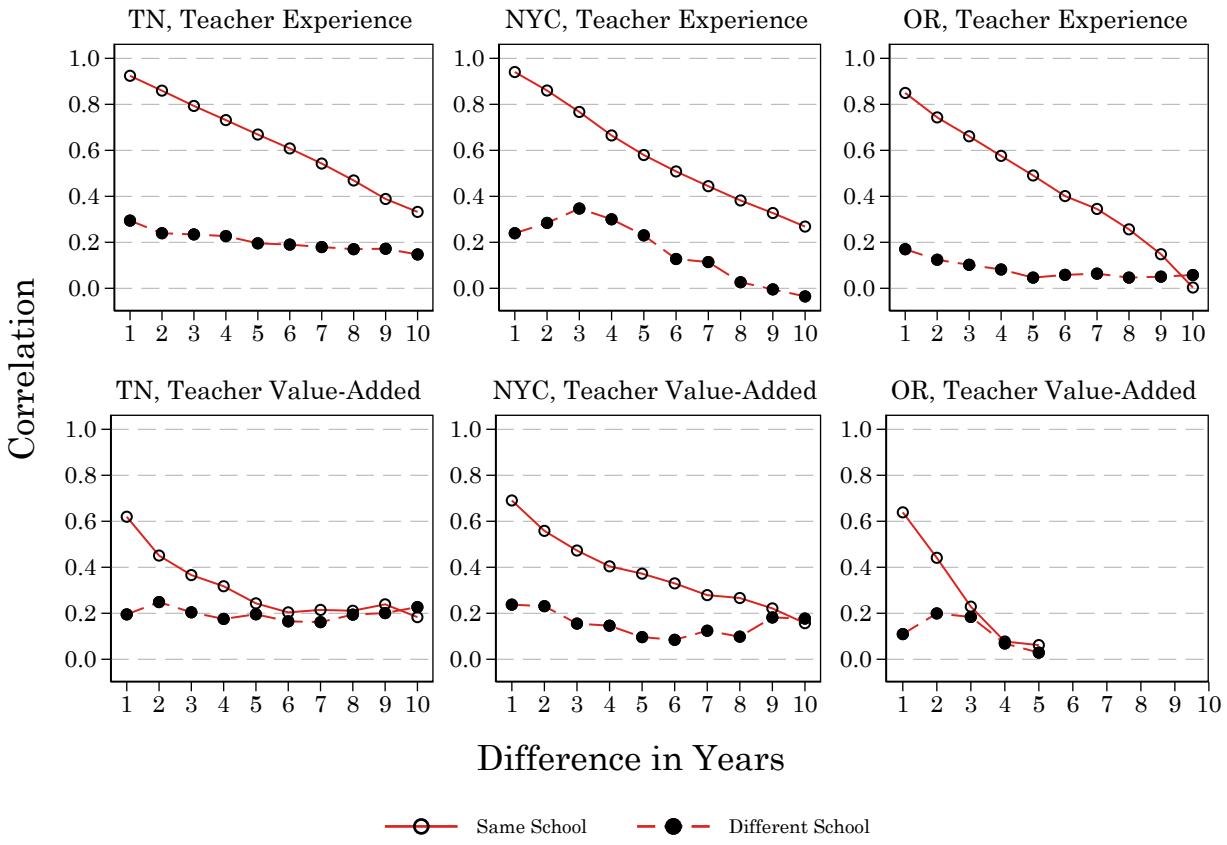


Figure F.2: Principal Autocorrelation Vectors for Teacher Composition

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.

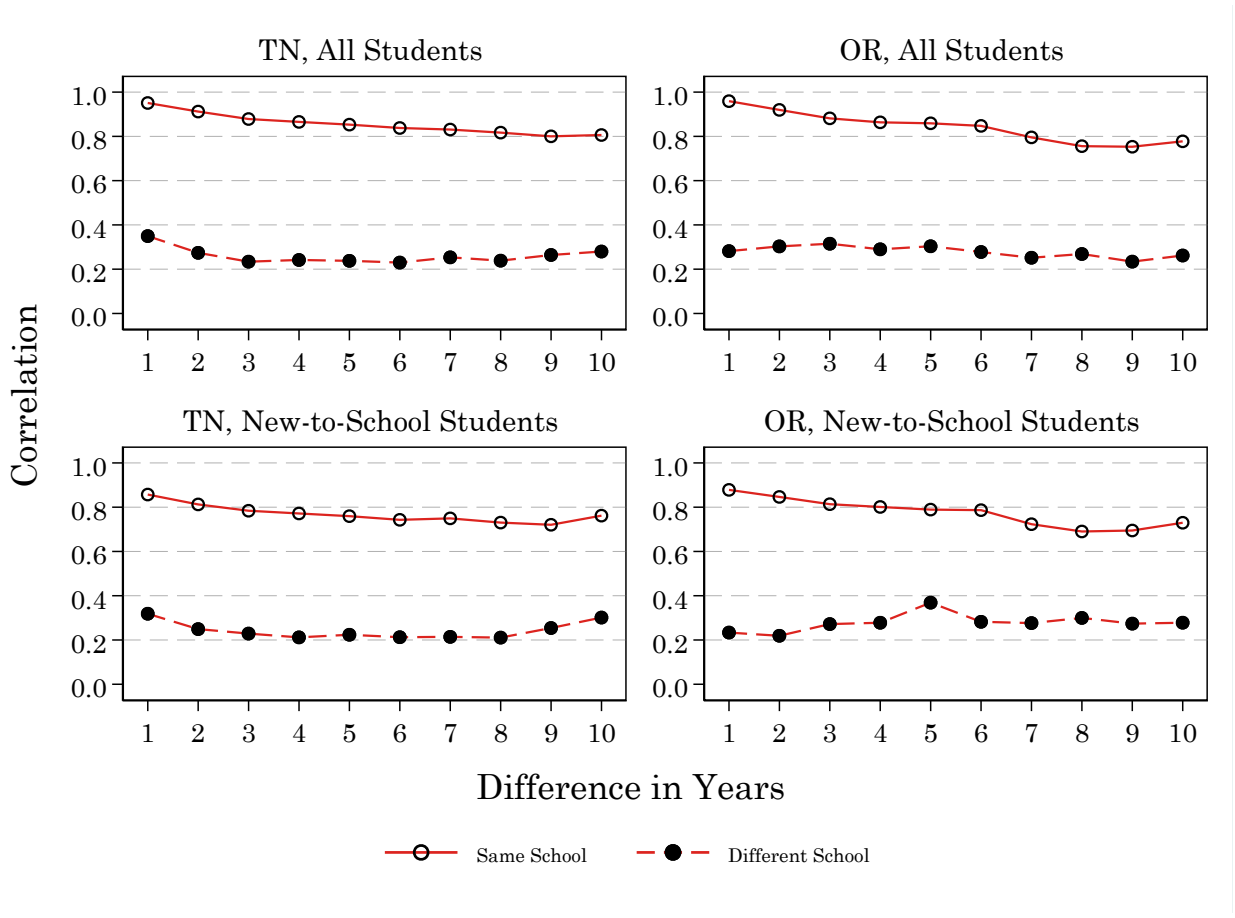


Figure F.3: Principal Autocorrelation Vectors for Student Composition

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 7. 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.

G Variance Decomposition with Autocorrelated Errors

Our primary analyses demonstrate that the between-principal variance component estimated in Table 1 largely reflects variation in school performance caused by factors outside of principals' control. Because these factors are partially persistent, they are not addressed through typical shrinkage approaches, such as Empirical Bayes's. Here, we demonstrate how a shrinkage approach that allows for autocorrelation in the school-by-year error term reaches a similar conclusion as our decomposition results based on $r^{\text{same}} - r^{\text{diff}}$.

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 illustrate how we can 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 8 (ϵ_{spt}) follows a second-order autoregressive structure:

$$\begin{aligned} \epsilon_{spt} &= \rho_1 \epsilon_{sp,t-1} + \rho_2 \epsilon_{sp,t-2} + \nu_{spt} \\ &\text{where } \nu_{spt} \sim N(0, \sigma^2) \end{aligned} \tag{13}$$

then, we can re-write Equation 8 with the composite error term:

$$\bar{Y}^*_{spt} = \theta_s + \theta_{s,p} + [\rho_1 \epsilon_{sp,t-1} + \rho_2 \epsilon_{sp,t-2} + \nu_{spt}] \tag{14}$$

where we directly estimate the AR(2) terms ρ_1 and ρ_2 along with the school, principal, and residual variance components. Note that we estimate an AR(2) model because we found that it fit better than an AR(1) model. 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 G.1. 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 AR(2) parameters confirm substantial positive autocorrelation in the residuals, which created the illusion of persistent between-principal variation in Table 1.

Table G.1: 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	0.128	0.175	0.162
Principal	0.000	0.000	0.000	0.000	0.000
Residual	0.216	0.213	0.179	0.194	0.188
AR(2) Parameters					
Correlation $(t-1)$	0.567	0.581	0.365	0.431	0.520
Correlation $(t-2)$	0.045	0.045	0.119	0.087	0.097
Panel B: New York City					
Random Effects Parameters (SD)					
School	0.499	0.235	0.114		
Principal	0.000	0.000	0.015		
Residual	0.222	0.186	0.126		
AR(2) Parameters					
Correlation $(t-1)$	0.744	0.667	0.273		
Correlation $(t-2)$	0.062	0.079	0.194		
Panel C: Oregon					
Random Effects Parameters (SD)					
School	0.362	0.321	0.170	0.243	0.264
Principal	0.000	0.000	0.035	0.000	0.000
Residual	0.183	0.177	0.152	0.176	0.158
AR(2) Parameters					
Correlation $(t-1)$	0.572	0.575	0.248	0.423	0.527
Correlation $(t-2)$	0.036	0.039	0.098	0.092	0.015

Notes: Cells report standard deviations of variance components and percentage of overall variance explained from Equation 14. 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 G.2: 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	0.293	0.119	0.168	0.143
Principal	0.000	0.000	0.020	0.000	0.000
Residual	0.148	0.140	0.106	0.125	0.115
AR(2) Parameters					
Correlation $(t-1)$	0.528	0.524	0.179	0.267	0.435
Correlation $(t-2)$	0.076	0.076	0.136	0.111	0.120
Panel B: New York City					
Random Effects Parameters (SD)					
School	0.486	0.239	0.110		
Principal	0.000	0.000	0.010		
Residual	0.194	0.157	0.108		
AR(2) Parameters					
Correlation $(t-1)$	0.670	0.586	0.172		
Correlation $(t-2)$	0.115	0.121	0.172		
Panel C: Oregon					
Random Effects Parameters (SD)					
School	0.339	0.288	0.153	0.220	0.202
Principal	0.000	0.000	0.000	0.035	0.000
Residual	0.159	0.152	0.138	0.148	0.132
AR(2) Parameters					
Correlation $(t-1)$	0.525	0.509	0.236	0.306	0.403
Correlation $(t-2)$	0.076	0.096	0.162	0.122	0.091

Notes: Cells report standard deviations of variance components and percentage of overall variance explained from Equation 14. 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.