ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 21-397

Connected Networks in Principal Value-Added Models

Brendan Bartanen
University of Virginia

Aliza N. Husain
Pivot Learning

A growing literature uses value-added (VA) models to quantify principals' contributions to improving student outcomes. Principal VA is typically estimated using a connected networks model that includes both principal and school fixed effects (FE) to isolate principal effectiveness from fixed school factors that principals cannot control. While conceptually appealing, high-dimensional FE regression models require sufficient variation to produce accurate VA estimates. Using simulation methods applied to administrative data from Tennessee and New York City, we show that limited mobility of principals among schools yields connected networks that are extremely sparse, where VA estimates are either highly localized or statistically unreliable. Employing a random effects shrinkage estimator, however, can alleviate estimation error to increase the reliability of principal VA.

VERSION: April 2022

Suggested citation: Bartanen, Brendan, and Aliza N. Husain. (2022). Connected Networks in Principal Value-Added Models. (EdWorkingPaper: 21-397). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/5tjj-py73

Connected Networks in Principal Value-Added Models

Brendan Bartanen[1] and Aliza N. Husain[2]

[1]University of Virginia

[2]Pivot Learning

April 2022

Author Note

Corresponding author: Brendan Bartanen, bartanen@virginia.edu. We are grateful to Dale Ballou, Jason Grissom, Matt Kraft, and Jim Wyckoff for their comments on the paper.

Abstract

A growing literature uses value-added (VA) models to quantify principals' contributions to improving student outcomes. Principal VA is typically estimated using a connected networks model that includes both principal and school fixed effects (FE) to isolate principal effectiveness from fixed school factors that principals cannot control. While conceptually appealing, high-dimensional FE regression models require sufficient variation to produce accurate VA estimates. Using simulation methods applied to administrative data from Tennessee and New York City, we show that limited mobility of principals among schools yields connected networks that are extremely sparse, where VA estimates are either highly localized or statistically unreliable. Employing a random effects shrinkage estimator, however, can alleviate estimation error to increase the reliability of principal VA.

Keywords: Value-added models, school leadership, principal quality, panel data methods

Connected Networks in Principal Value-Added Models

Introduction

A growing literature seeks to estimate *principal value-added* (VA): statistical models that isolate the contributions of individual principals to school performance, most often conceptualized as student test score gains, in an education production function. VA methods applied to principals can provide answers to two important questions: (1) Who is an effective principal? (2) How important are principals as inputs to student learning? Extant work consistently finds that principals matter, with the magnitude of principal effects typically ranging between 0.05 and 0.20 student-level standard deviations (SD) (Bartanen, 2020; Branch, Hanushek, & Rivkin, 2012; Chiang, Lipscomb, & Gill, 2016; Dhuey & Smith, 2018; Grissom, Kalogrides, & Loeb, 2015). In other words, a 1 SD increase in principal VA increases student achievement by 0.05 to 0.20 SD.

A key empirical challenge to estimating principal effects is to account for myriad school- or district-level factors that affect student learning but that principals cannot control. For example, principals cannot control the neighborhood in which the school is located and they similarly face constraints around teacher hiring and/or retention due to district policies or local labor market conditions. These school factors are often difficult to measure and may not be well-captured by available proxy measures, such as average student demographics. The approach taken in prior studies is to estimate a two-way regression model that includes principal and school fixed effects (FE), under the assumption that such unobserved school factors are largely fixed across time. In this model, identification of the principal fixed effects (i.e., the VA estimates) is restricted to within-school variation, but additional across-school comparisons of principals are possible if some principals work in multiple schools over time. Mobility groups formed by principals and schools due to principals transitioning across schools are termed “connected networks” in the principal VA literature (e.g., Bartanen, 2020; Burkhauser, 2017; Chiang et al.,

2016).[1]

The appeal of the connected networks model is that it can yield VA estimates where principals have a large comparison set, while also avoiding misattribution of school effects to principals. Despite its common use in estimating principal VA, however, our understanding of the properties of estimates from the connected networks approach remains limited. The inclusion of school FE creates additional challenges stemming from the limited mobility of principals among schools. Whereas other applications of two-way network models may benefit from observing many individuals working in multiple firms (or teachers in multiple schools), a school typically has only one principal at time and a majority of principals lead only one school in their career. This limited mobility potentially leads to weakly identified FE estimates that are unreliable measures of principal quality and overstate the magnitude of principals' effects (Jochmans & Weidner, 2019).

Connected networks models further hinge on the fundamental assumption that a principal's effectiveness is the same in any two schools. This assumption allows for the indirect comparison of principals who never worked in the same school, which is a key practical benefit of the connected networks model. Given prior work demonstrating that leadership is a relational process and that principals' impacts on student achievement are largely mediated through other school-level factors (e.g., Hallinger & Heck, 1998; Sebastian & Allensworth, 2012), such an assumption may be unrealistic. In particular, there may exist principal–school complementarities whereby part of a principal's effect reflects how well matched they are to a particular school context (Dhuey & Smith, 2018).

Using a simulation approach, this paper helps to fill a gap by investigating the accuracy of connected network models for estimating principal effects. Simulation studies have the advantage of creating controlled conditions where the performance of VA estimators can be tested according to different assumptions about the data-generation

[1] While defined later, a "connected network" comprises the largest possible set of schools in which every school has had at least one principal move to at least one other school in the network.

process. In particular, this approach allows us to focus on the aforementioned issues related to the connected networks approach. While we acknowledge that principal VA modeling faces additional challenges beyond those we examine here, such as the assumption that a principal's impact is immediate and fixed over time, our insights here contribute to a broader effort to improve methods for measuring principal quality using student outcome data. In that sense, our simulation is conceptually similar to those used to examine the accuracy of teacher VA models under different assumptions about the nonrandom sorting of students to teachers (e.g., Guarino, Maxfield, Reckase, Thompson, & Wooldridge, 2015; Guarino, Reckase, & Wooldridge, 2015). To supplement the insights drawn from our simulation, we also provide a brief empirical application using actual test scores.

Our simulation is built both from administrative datasets from Tennessee and New York City (NYC). That is, using the connected networks formed by actual mobility patterns of principals and schools over long panels, we generate simulated test scores where the true principal effects are known. We then apply VA models to the simulated data and compare the estimated and true principal effects. To more deeply understand the labor market dynamics that lead to these networks, we also construct datasets from a simulated process of job separations that varies the degree of switching of principals among schools. We consider two questions regarding the performance of principal VAMs: (1) How accurately do VAMs rank principals according to their true effects? (2) To what extent does the magnitude of principal VA accurately reflect the true magnitude of principals' effects? Answers to these questions can provide insight about whether principal VA models are likely to provide accurate results in real-world conditions.

Our results uncover a key tradeoff between the statistical precision of principal VA estimates and their practical utility. Even in large-scale datasets where we observe thousands of principal transitions, the underlying network structure of principals and schools is very weak because most turnover events are exits from the principalship rather than across-school transfers. This yields two distinct types of connected networks. First,

many principals belong to small networks that contain one or two schools, meaning that their estimated VA reflects performance relative to only a handful of other principals. While principal VA from small connected networks is precisely estimated, such localized performance measures may lack practical usefulness (e.g., as an accountability metric). Principals in large networks, on the other hand, can be compared to hundreds of other principals. However, the underlying network is weakly connected, undermining the reliability of VA estimates and producing inaccurate rankings of principals. Further, because schools in large connected networks are typically linked through only one or two mobile principals, inaccuracy is amplified substantially in the presence of principal–school complementarities.

A similar tradeoff exists for using principal VA models to understand the magnitude of principals' impacts on student outcomes. In small networks, school FE erroneously eliminate part of the real difference in principal quality, leading to an understatement of the importance of principals. While large networks circumvent this problem, they *overstate* the magnitude of principal effects because of the estimation error introduced by weak network structures.

Given the estimation error of principal VA estimates in large networks, we further examine whether shrinkage approaches can improve correlations between principals' estimated and true effects. Employing a mixed model that treats principal effects as random greatly reduces estimation error in large networks. This method improves the precision of principal VA and yields a substantially more accurate estimate of the magnitude of principal effects.

This paper contributes to our understanding of principal VA models, where evidence on their validity and reliability remains limited. This dearth of evidence stands in stark contrast to the teacher VA literature, where a large number of studies have investigated these properties (see Koedel, Mihaly, & Rockoff, 2015, for a review). We also contribute to a larger literature utilizing two-way regression models in the context of matched

employer–employee datasets. Most notably, we provide an application of recent theoretical work (e.g., Jochmans & Weidner, 2019; Kline, Saggio, & Sølvsten, 2018; Verdier, 2018) concerning inference of fixed effects estimated from network data. We also build on related applications using school fixed effects to control for unobserved school heterogeneity in value-added modeling, including for teachers (Mansfield, 2015) and teacher preparation programs (e.g., Mihaly, McCaffrey, Sass, & Lockwood, 2013). These applications highlight some key challenges for analyzing connected networks that we consider in the context of the principal labor market, with the added benefit of a simulation analysis that can provide deeper insight around the accuracy of VA models with school FE.

Conceptual Framework for Estimating Principal Value-Added

We begin our discussion of estimating principal VA by placing principal effects in an education production function, where achievement in a given grade is a function of school and non-school (i.e., student or family) inputs. As we outline in the remainder of the section, a principal's contribution to student achievement comes through their ability to improve the school-level inputs to which students are exposed. Given the nature of principal effects, a key challenge is to successfully isolate the principal's effect from the effect of school-level factors that the principal cannot control.

Education Production Function

Following prior work in the teacher effects literature (e.g., Guarino, Reckase, & Wooldridge, 2015; Sass, Semykina, & Harris, 2014), we conceptualize educational production using a cumulative effects model:

$$A_{it} = f_t(E_{it}, ..., E_{i0}, X_{it}, ..., X_{i0}, c_i, u_{it}) \tag{1}$$

where achievement for student i in grade t (A_{it}) is a function of time-varying school (E) and non-school (X) inputs, an unobserved and time-invariant student effect (c), and

idiosyncratic shocks (u). Note that (1) includes the full history of inputs from grade t to their first year of schooling (i.e., $t = 0$). The functional form f_t may vary across grades. Conceptually, a principal's contribution to A_{it} is through their ability to improve the school inputs to which students are exposed. For example, effective principals may strategically retain high-quality teachers, establish a positive school climate, and determine student-to-teacher assignments that improve student learning. These are improvements to E_{it}. As we discuss further below, there are elements of E_{it} that principals control minimally or not at all, such as the per-pupil funding level.

Given its general form, many of the potential elements of equation 1 are unobserved (e.g., family inputs to their child's schooling) and there is limited information regarding their relationships to one another—e.g., interactions, feedback loops, functional form (Guarino, Reckase, & Wooldridge, 2015). Therefore, moving from (1) to a model that is empirically tractable requires a number of simplifying assumptions. Typically, studies in the value-added literature assume a grade-invariant functional form that is both linear in parameters and imposes additive separability, which allows us to rewrite (1) as:

$$A_{it} = E_{it} + E_{i,t-1} + ... + E_{i0} + X_{it} + X_{i,t-1} + ... + X_{i0} + c_i + u_{it} \tag{2}$$

While (2) moves closer to a empirically tractable model, it remains infeasible given that typical administrative datasets do not contain much information about inputs included in E or X, particularly prior inputs. As a final simplification, we can use each student's prior-year achievement as a sufficient statistic for the history of prior inputs (for both E and X) they received under the assumption that these inputs decay geometrically with time. This reduces equation 2 to:

$$A_{it} = \lambda A_{i,t-1} + \beta E_{it} + \gamma X_{it} + \pi c_i + e_{it} \tag{3}$$

(3) is more parsimonious because it includes only a single lag of achievement ($A_{i,t-1}$) and

current-year school and non-school inputs (E_{it} and X_{it}). The parameter λ determines how quickly past inputs decay in terms of their impact on contemporaneous achievement. Student heterogeneity (c_i) remains in the model. E_{it} is typically reduced to a set of teacher indicators, whose coefficients β are the parameters of interest (teacher VA). For teacher VA, a key consideration is how to best account for nonrandom sorting of students to classrooms/teachers, typically on the basis of $A_{i,t-1}$ and/or c_i. Because c_i is unobserved, leaving it in the error term creates the potential for biased teacher effect estimates. However, both simulation and empirical evidence suggest that the best approach is to simply estimate (3) without directly accounting for c_i. This is because $A_{i,t-1}$ and c_i are almost certainly strongly correlated. While leaving c_i in the error term creates an upward bias in the estimated persistence parameter λ (Andrabi, Das, Khwaja, & Zajonc, 2011; Guarino, Reckase, & Wooldridge, 2015), this is a second-order concern given the focus on obtaining useful estimates of β. We return to this issue below when discussing the typical model for estimating principal effects.

The Principal's Contribution to School Performance

Shifting to the goal of estimating principal effects, we use (3) as a starting point given its parsimony and widespread application for value-added modeling. As noted previously, a principal's contribution to A_{it} comes through their ability to improve E_{it}—the school inputs that students receive. We thus might be tempted to estimate principal VA similar to teachers in conceptualizing E_{it} as a set of indicator variables for principals. However, E_{it} likely includes many variables—such as the effectiveness of teachers, the quality of school facilities, and school climate—over which principals have limited control. For instance, principals do not control the per-pupil funding formula or teacher salary schedule, which undoubtedly constrains their choice set regarding class size/assignment policy and teacher hiring. Further, many of the relevant school inputs are difficult to measure in typical administrative data and may only be weakly correlated with observables

like average student demographics. A key challenge, then, in estimating principal VA is to separate the principal's effect from school-level factors that she cannot control.

To make this discussion more concrete, we define θ_{jst} as an index summarizing the average quality of inputs E that school s with principal j provides to students in year t. We refer to θ_{jst} as *school performance*, which is a function of principal effectiveness (P) and aspects of school quality that principals cannot control (S):

$$\theta_{jst} = f(P_{j(t,s)}, S_s) \tag{4}$$

Note that the functional form of (4) is general. In the simplest formulation, we could assume principals and schools have effects that are additive and constant: $P_{j(t,s)} + S_s$. In fact, this is the implicit framework used in most prior studies. However, school performance may also be a function of the quality of the match between the principal and the school context.[2] In this case, a reasonable realization of (4) is:

$$\theta_{jst} = P_{j(t,s)} + S_s + M_{js(t,s)} \tag{5}$$

where M is the match quality between each principal-school pair, which is orthogonal to P and S. M could capture, for instance, a principal's leadership style that fits well with existing school staff. While (5) is more flexible in allowing for principal-school complementarities, it maintains the assumption that P, S, and M are fixed over time.

The goal of principal VA modeling is to successfully isolate P or $P + M$ from S. That is, we want to measure principal effectiveness in a way that avoids attributing to principals factors that they cannot control. This is conceptually similar to teacher VA,

[2] While there has been little empirical work examining principal–school complementarities in the context of VA models (see Dhuey & Smith, 2018, for an exception), prior studies demonstrate that principals operate within the contexts of their schools. For instance, research shows that Black principals increase the likelihood of Black teachers being hired and retained (Bartanen & Grissom, 2021), and male teachers are more likely to turn over under female principals and request to transfer to schools with male principals (Husain, Matsa, & Miller, 2018). These studies suggest that principals experience varied levels of success based on teacher demographics in their schools, supporting the potential for match effects.

where the chief concern is to avoid punishing or rewarding teachers based on the pre-existing characteristics of their assigned students. Whether the estimand is P or $P + M$ (or P and M separately) depends on the intended use. While there are certain scenarios where isolating the fixed component of principal quality is desirable (e.g., wanting to understand whether a principal is likely to be effective in a different school), our primary aim is to evaluate models that attempt to measure the principal's overall contribution $(P + M)$ to improving student achievement in the schools where they actually worked, regardless of whether it reflects a fixed or match effect. This also greatly reduces the number of parameters to be estimated, which is important given the limited mobility of principals among schools, which we discuss in depth below.

The Principal and School Fixed Effects Model

As discussed above, a key challenge to accurately measuring principal effectiveness is to avoid misattributing to principals the fixed characteristics of their school. The nearly universal approach in extant work is to estimate a model with principal and school fixed effects (FE), where the

$$A_{ijst} = \lambda A_{i,t-1} + \eta \mathbf{X}_{ijst} + \delta_j + \gamma_s + e_{ijst} \tag{6}$$

where the principal FE (δ_j) represents value-added. On a fundamental level, a principal's VA estimate in (6) is the mean test score residual across all of the students who were in the school during the principal's tenure. A high-VA principal, then, is one whose students tend to have positive test score residuals. Crucial to this approach is that the residualization adequately accounts for factors that should not be attributed to principal quality, such as student background characteristics and fixed school factors. Thus, principal VA models typically control for student characteristics (race/ethnicity, gender, indicator for free/reduced-price lunch eligibility, etc.) and school-by-year aggregates of these student characteristics. These are represented by $\mathbf{X}_{ijst}$ in (6). Along with $A_{i,t-1}$ and $\mathbf{X}_{ijst}$, the

school FE γ_s aim to control for myriad factors that potentially affect current-year student achievement but should not be credited to principals.

The school FE are critical because even with controls for observable student and school characteristics, there likely remains substantial unexplained school-level variation. The quality of the school's neighborhood, for instance, may only be partially captured by students' background characteristics.[3] The key identification assumption for (6) to yield unbiased estimates of principals' effects is that any unobserved school-level heterogeneity is fixed over time. Any unobserved time-varying factors will become part of δ_j, introducing bias and potentially leading to an overstatement of the magnitude of principals' effects. This is a strong assumption that has received inadequate attention in the extant literature, but we maintain it here to maintain our focus on the connected networks problem. An additional assumption, which we unpack below, is that the matrix described by (6) is full rank. That is, estimating both δ_j and γ_s requires sufficient variation to avoid perfect multicollinearity.

In the case of teacher VA, models typically do not include school FE and students have new teachers each year. Thus, the teacher effect is not a structural component of $Y_{ijs,t-1}$. In the case of principals, both the school and principal contribute to the prior-year score. Any bias in λ, then, will necessarily lead to bias in principal and school effects. However, because principals and schools have *continued* impacts on student achievement and the prior-score is a good proxy for unobserved student heterogeneity, the bias does not substantially diminish the accuracy of the principal VA estimates.

To summarize, principal VA models attempt to isolate variation in students' test score growth that is attributable to the principal's effectiveness as opposed to factors that

[3] We provide empirical support for this claim in Appendix Table B1, which shows variance component estimates from mixed models in Tennessee and New York City using different sets of controls. Even after controlling for prior-year test scores, student characteristics, and school-by-year means of student characteristics, the school-level variance component remains roughly equal in magnitude to the principal variance component, underscoring the potential for bias in principal VA estimates that do not account for unobserved school heterogeneity.

she cannot control. Given the indirect nature of principals' effects—through shaping school-level factors such as climate and human capital as opposed to direct instruction in classrooms—it is critical to account for unobserved school-level heterogeneity. Nearly all existing studies estimate models that include school FE, which controls for any fixed differences between schools. We now turn to the conceptual and practical challenges of this approach.

Connected Networks in Principal Value-Added Models

Estimating effects for both principal and school in (6) is challenging because each school has only one principal at a time, meaning that in cross-sectional data δ_j and γ_s are perfectly collinear. With panel data of sufficient length, however, schools will have multiple principals, which creates the necessary within-school variation required to estimate coefficients for δ_j. The interpretation of principal VA estimates between models with and without school fixed effects, however, can be very different. Without school fixed effects, VA estimates produce a global ranking of all observed principals. When school FE are included, principal VA estimates often produce *local* rankings. Specifically, with school FE, principals can only be compared within a *connected network* of schools, where a network is the largest possible set of schools in which every school has had at least one principal transfer to at least one other school in the network during the analysis period.

The size of a connected network can range from a single school to the entire set of schools, depending on the number of schools and years in the panel and the mobility patterns of principals across schools. A single-school network will result when none of the principals who worked in that school were observed working in a different school. A multi-school connected network will form when a principal who works in school A moves to school B. This connection allows for the comparison of all principals who ever worked in school A or school B. If school A or school B further has a principal who also worked in school C, the connected network grows to include all principals who ever worked in one of

these three schools. Given sufficient mobility, a connected network can theoretically include the complete set of observed schools. In this case, estimates of δ_j in 6 would yield a global ranking of principals.

In practical applications, however, connected networks of principals and schools tend to be small, often comprised by only a single school. Even when longer panels of data are available, high principal attrition rates (i.e., exiting from the principalship entirely) mean that there are relatively few principals who transfer between schools.[4] Nationally, roughly 20% of principals leave their positions each year, but two-thirds of this turnover is movement out of a principal position (retirements, moving to a central office position, etc.) (Grissom et al., 2019).[5]

While the interpretation of principal VA estimates in network models as local measures of principal effectiveness has been well-established in prior studies, there has been virtually no work that investigates their validity and reliability. Consequently, we know little about whether these models produce accurate measures of principals' effects on student outcomes. While the inclusion of school FE is conceptually appealing as a means to control for unobserved factors, it also introduces a number of additional challenges that have not been given much attention. Our analysis focuses on three of these challenges, which we outline below.

[4] Additionally, because most datasets encompass a single district or state, some real movement across schools will be missed, worsening this problem. However, this isn't likely to be a major issue, as prior evidence suggests that across-district movement of principals is relatively rare. (e.g., Grissom, Bartanen, & Mitani, 2019)

[5] Unsurprisingly, then, state-level applications of principal VA find small network sizes. For instance, when examining data for Pennsylvania students in grades 4–8 from the 2008–09 to 2012–13 school years (a comparatively small panel), Chiang et al. (2016) find that 76 percent of connected networks are single-school and only 2.6% of networks included four or more schools. Similarly, of the connected networks present in Burkhauser's (2017) data spanning 2005–06 to 2011–12, 57% were networks with only two principals, indicating that a majority of school leaders were being compared to only one other leader. In a statewide panel spanning 10 years, Bartanen (2020) observes almost 20% of principals in networks with 10 or more schools, with 39% of principals in single-school networks.

Estimation Error in Sparse Networks

While prior studies make clear that including school FE in principal VA models changes the interpretation of a principal's estimated effect, they fail to note that school FE exacerbates estimation error, which lowers reliability and leads to upward bias in the estimated magnitude of principal effects. Recently, econometricians have paid increasing attention to two-way regression models using network data (Jochmans & Weidner, 2019; Verdier, 2018); the principal and school fixed effect model is one application of such models. When examining these models, Jochmans and Weidner (2019) demonstrate that the statistical precision of individual effects is determined by the connectivity structure of the underlying network. In the case of principal VA, limited mobility of principals among schools means both that many principals are in small networks and that estimates for principals in larger networks contain considerable noise. Intuitively, this noise comes from variance inflation, as indicator variables for each principal and school are highly correlated with one another.

Specifically, Jochmans and Weidner (2019) show that the two-way fixed effects model (in our case, principals and schools) can be analyzed as a weighted bipartite graph, where edges connecting principals to schools are weighted by the number student-by-year observations. The statistical precision of the principal fixed effects is determined by how strongly connected principals are within the given connected network. In particular, they demonstrate that bottlenecks in the network (i.e., where two larger sets of principals are connected only through a single principal) lead to variance inflation and, ultimately, imprecise VA estimates. Mathematically, these bottlenecks can be summarized by the smallest nonzero eigenvalue (λ_2) of the graph's normalized Laplacian matrix, where $\lambda_2 \to 0$ as the network becomes more sparse.

To make more concrete the concept of network connectivity, Figure 1 shows two examples of medium-sized principal networks in Tennessee. In each plot, the numbered nodes represent principals, with edges representing comparisons among principals who

worked in the same school. As an example, the left part of plot A shows that principals 2, 3, and 11 worked in the same school and can be directly compared. Principal 3 also worked in a (different) school with principals 7 and 16. This allows for indirect comparisons between principals in these subsets. As more principals switch schools (as opposed to leaving the principalship altogether), networks will grow larger and more indirect comparisons will become possible.

While the networks in plots A and B are similar in the number of principals and schools, their connectivity (summarized by λ_2, the smallest non-zero eigenvalue) differs. Intuitively, a network is weakly connected when it is easy to separate it into two substantial sub-networks by removing edges. This is the case for plot A, where there are fairly few edges linking the two sides, and these centralized edges have relatively low weights (denoted in the plot by the edge width). In plot B, there are far more edges, which represents greater mobility among schools. There are also redundancies such that the network cannot be split into two sub-networks by removing a single principal. As a result, variance inflation in the network shown in plot A is predicted to be roughly four times greater than in the network shown in plot B.

Variance inflation in principal VA models has not been formally investigated, though prior work tends to conclude that VA estimates are precise, given the large number of student-by-year observations that contribute to estimating each principal's effect. Thus, our analysis contributes by examining variance inflation due to the structure of connected networks, which we show theoretically using the techniques from Jochmans and Weidner (2019) and through our simulation that uses the actual connected networks of principals in Tennessee and New York City, as well as networks formed by simulated labor markets, where we can indirectly manipulate connectivity and network size.

Downward Bias of the Variance in Small Networks

While estimation error due to sparse network structure creates an upward bias in the estimated magnitude of principal effects, an additional challenge introduced by the inclusion of school fixed effects is a *downward* bias in the estimated magnitude, concentrated in small networks that have few principals. To see this, consider a scenario where each school has two principals and no principals switch schools, thus making each school its own connected network. In expectation, the mean principal effect in each network is zero, but the observed mean will be nonzero due to sampling variation. In the connected networks approach, this sampling variation—which reflects real information about principal quality—will be captured by the school fixed effect, which creates a downward bias in the estimated variance of the principal effect that decreases as the number of principals increases. This downward bias works in the opposite direction as estimation error from the sparse network structure.

As connected networks grow larger, this downward bias in the variance will decrease, which will make the empirical distribution of VA estimates a better representation of the true variance of principal effects. As described in the previous section, however, large networks may suffer from greater estimation error.[6]

Principal–School Complementarities

As outlined in the connected networks section, principal VA models that include school FE result in VA estimates that are local to the principal's connected network. While the identifying variation is restricted to comparisons of principals who worked in the same school, indirect comparisons are made possible via principals who work in multiple schools

[6] These dynamics may help to explain why prior studies reach somewhat different estimates of the magnitude of principal effects despite similar empirical approaches. Bartanen (2020) and Dhuey and Smith (2018), for instance, draw on long panels from statewide data (where more principals are in large networks) and find larger SD of principal VA estimates. By contrast, Grissom et al. (2015) and Branch et al. (2012) find lower SDs. In the former study, principal VA is estimated using just a single district across an 8-year panel. While Branch et al. (2012) draw on statewide data from Texas, they restrict the size of connected networks to a single school by estimating principal-by-school FE.

across the study period. Intuitively, principal A, who worked in school X, can be compared to principal B, who worked in school Y, if there exists a principal C who worked in both schools. Connected networks can grow large as the length of the panel increases or as principals move between schools more frequently.

Underpinning these indirect comparisons of principals who never worked in the same school is the assumption that the effectiveness of the mobile principal (i.e., the one who connects the two schools) is fixed. More specifically, the connected networks approach requires an assumption that there are no complementarities or "match effects" between principals and schools. Taking the simple example above, comparing principal A in school X to principal B in school Y breaks down if principal C's effectiveness is different in school X versus school Y. If, for example, principal A and principal B are equally effective, but principal C is better matched in school X than school Y, then principal A will appear less effective than principal B in the connected networks model.

At first blush, incorporating match effects into a principal VA model seems nearly impossible given that the vast majority of principals are not observed in multiple schools. As our prior results demonstrate, even separating fixed principal and school effects places considerable constraints on the data. Match effects add yet another layer of complexity. Further, without considerable principal mobility, FE strategies to identify match effects will suffer from a considerable small sample bias (Jackson, 2013). Nonetheless, we can examine how the accuracy of principal VA models—which typically assume portability of principal effectiveness across schools—changes when match effects are a large component of the principal effect. We also examine match effects in a random effects framework, which is described in the modeling section.

Simulation

To examine the accuracy of principal VA from connected network models, we employ a simulation that compares principals' VA estimates to known effects, which are drawn

randomly. Simulation approaches have been used previously to examine teacher VA (e.g., Guarino, Maxfield, et al., 2015; Guarino, Reckase, & Wooldridge, 2015). Different from simulation studies of teacher VA, however, we use both simulated data and administrative data from Tennessee and NYC as the structure of the simulation. For our simulated labor market, which is described in detail in Appendix C, we construct principal-to-school datasets through a job separation and clearing process that simulates the formation of connected networks. This is important because the focus of our study is the connected network approach, and the accuracy of these estimates depends on the network structure of the dataset. For the TN and NYC networks, we start with a dataset at the student-by-year level that contains unique identifiers for student, principal, and school. We then randomly draw the principal and school effects, and generate student outcomes as a function of these effects. Below, we outline our assumed data-generation process and simulation procedures.

Data-Generating Process

To isolate issues relevant to the connected networks approach, we assume a fairly straightforward data-generating process for student test scores:

$$A_{ijst} = \lambda A_{ijs,t-1} + \theta_{jst} + c_i + e_{ijst} \tag{7}$$

where i, j, s, and t index student, principal, school, and year, respectively. $A_{ijs,t-1}$ is the prior-year score with a persistence parameter λ, θ_{jst} is the school-by-year-specific contribution to the current-year score, c_i is time-invariant student heterogeneity, and e_{ijst} is a random error term that is assumed to be independent over time. This mirrors the DGP used by Guarino, Maxfield, et al. (2015); Guarino, Reckase, and Wooldridge (2015) in their teacher VA simulations, except that we conceptualize school-level inputs as a single school-by-year effect rather than a set of teacher indicator variables. We refer to the school-by-year effect as "school performance," which is a linear function of fixed principal quality (δ_j), a principal-by-school match effect (α_{js}), fixed school-level factors that the

principal cannot control (γ_s), and a school-by-year random shock (v_{jst}):

$$\theta_{jst} = \delta_{j(t,s)} + \alpha_{js(t,s)} + \gamma_s + v_{jst} \tag{8}$$

In this DGP, changes in school performance (aside from yearly random deviations) are completely determined by the principal.

As with any simulation approach, we acknowledge that this DGP is a simplification, and that the true nature of principal effects may be substantially more complex. In particular, we are assuming that principal and school quality are fixed and that unobserved student heterogeneity has a constant effect in each year. We additionally assume no time-varying student or family effects, no interactions between students and principals or schools, and no peer effects. Finally, we assume that test scores have no measurement error and there is no serial correlation in the error term. These simplifications allow us to understand more deeply how the principal and school fixed effects model may or may not produce good estimates of principal quality according to the structure of connected networks. If the models perform poorly here, they likely face greater challenges in the context of real data.

We show our simulation parameters in Table 1. Panel B shows the parameters that are fixed across all simulations, which we chose following Guarino, Reckase, and Wooldridge (2015). Specifically, we assume a persistence parameter of 0.5, implying that past school and family inputs decay geometrically across years (Sass et al., 2014), though our results are not particularly sensitive to this choice. For the magnitude of the school performance effect, we assume schools are responsible for 5% of the total variance in student achievement growth, which corresponds roughly to the lower end of the range found using variance decomposition methods for math and reading scores in Tennessee and New York City.[7]

[7] These results are shown in Appendix Table B1. Specifically, we estimate a school-by-year random effects model for current-year test scores with controls for prior-year test scores, student characteristics, and

Panel C shows the parameters that we vary across simulations: (1) the relative magnitude of the principal-by-school match effect and (2) the correlation between principal quality and the fixed school effect. Across the simulations, we hold constant the relative importance of the principal (45% of the variance of the school performance effect), school (45%), and random shock (10%), but we vary how much of the principal effect is the fixed component (δ_j) versus the match component (α_{js}).[8] Specifically, we test models where there is no match effect and where the match effect is slightly larger than the stable principal effect. We also explored an intermediate case with a small match effect, but omit those results for parsimony.[9] Second, we examine different correlations (0.4, 0, and -0.4) between the fixed principal and school effects, where a positive correlation means that effective principals are more likely to work in effective schools.

Models for Estimating Principal VA

The purpose of our simulation is to compare principals' true effects to their estimated effects from VA models using principal and school fixed effects. Specifically, we estimate:

$$A_{ijst} = \tilde{\lambda} A_{ijs,t-1} + \tilde{\delta_j} + \tilde{\gamma_s} + e_{ijst} \quad (9)$$

As previously described, this model produces estimates of principal effects ($\tilde{\delta_j}$) that are relative to the mean of principals within the same connected network. We refer to this model as "principal FE + school FE" (P+S FE).[10]

---

school-by-year means of student characteristics.

[8] Supporting this choice, in the variance decomposition results shown in Appendix Table B1, we find that the variance components for principals and schools are roughly equal for both math and reading.

[9] We found the results were always bounded within the two extreme cases and thus provided no additional insight about the dynamics.

[10] One challenge is the estimation of $\tilde{\lambda}$, which is the persistence parameter for prior-year test score, $Y_{ijs,t-1}$. As noted in prior work, this estimate is biased upwards due to the presence of α (fixed student heterogeneity) in the error term (Andrabi et al., 2011; Guarino, Reckase, & Wooldridge, 2015). Further,

In addition to the P+S FE model, We also examine several alternative specifications. First, to understand the importance of variance inflation introduced by the sparse networks of principals and schools, we estimate models that restrict the size of these networks. Specifically, we replace principal FE with principal-by-school FE, which we refer to as "principal-school + school FE" (P-S+S FE). By estimating an effect for each principal-by-school spell rather than each principal, connected networks are restricted to a single school—principals cannot be compared across schools. While this greatly limits the comparison set for many principals, it also reduces the noise component inherent to the P+S FE model.

Instead of reducing the size of networks, an alternative to potentially reduce the variability of principal VA estimates is to implement a shrinkage estimator, such as Empirical Bayes's (EB), that adjusts for the estimation error in the principal FE. The intuition of the EB approach, in this case, is that principals with very high (low) VA estimates are likely suffering from positive (negative) estimation error. EB estimation accordingly shrinks these estimates toward the mean principal effect, yielding a biased but less noisy VA estimate. In theory, the shrunken estimates should have a higher correlation with the true principal effects.

Our preferred approach to implementing EB is to estimate a mixed model where principals and schools are random effects instead of fixed effects (P+S RE). From this model we obtain the best linear unbiased predictions (BLUPs) for the principal effects. We also examine a second procedure whereby we make a post hoc adjustment to the estimated FE by drawing on their standard errors as a measure estimation error, according to the

---

because most students remain in the same school (and have the same principal) between year $t-1$ and t, the upward bias in $\tilde{\lambda}$ leads to attenuation of $\tilde{\delta}$ and $\tilde{\gamma}$. Consistent with prior studies, however, we find that the impact of this bias on the accuracy of VA estimates is small, and we thus proceed with the lagged score model, which is the common approach for estimating principal effects.

following formula:

$$\hat{\delta}_j^{EB} = \lambda_j \hat{\delta}_j^{FE} + (1 - \lambda_j)\bar{\delta} \tag{10}$$

The FE estimate $\hat{\delta}_j^{FE}$ is shrunken towards the mean principal effect ($\bar{\delta} = 0$) by the factor $1 - \lambda_j = 1 - \frac{\hat{\sigma}_\delta^2}{\hat{\sigma}_\delta^2 + \hat{\zeta}_j}$. This shrinkage factor is a function of the estimated variance of principal effects ($\hat{\sigma}_\delta^2$) and the estimated error variance of principal j's effect. The latter quantity is simply the squared standard error of the FE estimate for principal j, while the former is approximated by the mean of the square of the standard errors of $\hat{\delta}_j$ subtracted from the variance of $\hat{\delta}_j$ (e.g., Aaronson, Barrow, & Sander, 2007; Branch et al., 2012).[11] Intuitively, as error in a principal's FE estimate ($\hat{\zeta}_j$) increases relative to the variance of principal effects ($\hat{\sigma}_\delta^2$), the shrinkage factor ($1 - \lambda_j$) increases, pulling the EB estimates towards zero. Our main results focus on the random effects approach given the extreme computational demands of obtaining standard errors for the shrunken FE, though we compare these approaches using a subset of the data in Appendix D. We also implement a version of the random effects approach that includes an explicit match component: principal RE + school RE + principal-school RE (Woodcock, 2015).

Finally, we estimate a model that does not include school FE, which we call "principal FE only" (P FE). This is effectively school value-added averaged over the principal's tenure in the school. While we anticipate that this model will perform poorly in scenarios where the fixed school component is a large contributor to school performance, omitting school FE avoids the estimation challenges endemic to the connected networks approach and allows for a global ranking of principals. In particular, this models helps us to understand whether the bias/precision tradeoff makes sense when choosing to include school fixed effects in principal VA models.

[11] We obtain the standard errors for the FE using the routine proposed by Mihaly, McCaffrey, Lockwood, and Sass (2010), which accounts for the sum-to-zero constraints within connected networks.

Assessing Model Performance

For each of the six unique combinations of the simulation parameters in Table 1, we run 25 Monte Carlo replications of the simulation, where the principal, match, school, and student effects are drawn randomly. Given the large-scale nature of our datasets, our results are highly consistent across replications. In the performance metrics described below, we report the simple average across the 25 replications.

We consider two main questions about the performance of principal VA models. (1) How accurately do VAMs rank principals according to their true effects? (2) To what extent does the magnitude of principal VA accurately reflect the true magnitude of principals' effects? To answer these questions, we draw on simulated data to compare the VA estimates to the true principal effects. In the models that contain school FE, however, we must first adjust the true principal effects by centering them within connected networks.[12] Importantly, we define the true principal effect to include both the fixed and match components of the principal effect.[13]

We then report three summary measures. First, we compute the Pearson correlation between the principal VA estimates and the true principal effects.[14] A correlation of one, to be specific, would indicate that the model perfectly ranks principals in terms of their true effectiveness (within networks). Low correlations between estimated and true principal effects may be a product of bias, imprecision, or both. Thus, we also report performance measures that isolate these factors. Our second measure captures bias in the VA estimates

[12] Specifically, we residualize true principal effects on the vector of network FE. For the P-S+S approach, we effectively treat each principal-by-school spell as a separate principal, though our results are essentially identical if we weight our performance metrics inversely by the number of spells per principal.

[13] For principals who work in multiple schools, we construct a time-invariant total principal effect that uses a weighted average (according to the number of student-by-year observations) of the match components from each of their schools.

[14] Using Spearman rank correlations yields very similar estimates.

by estimating a simple regression of VA estimates as a function of true effectiveness:

$$\tilde{\delta_j} = \beta\delta_j + e_j \tag{11}$$

where $\tilde{\delta_j}$ is principal j's VA estimate and δ_j is their true effect. $\beta = 1$ would indicate that the VA model produces unbiased estimates of principals' true effects. If $\beta > 1$ ($\beta < 1$), the VA estimates amplify (condense) the true principal effects. A model that produces VA estimates where $\beta > 1$ may correctly rank principals even though the estimates are systematically biased. Note that because $\tilde{\delta_j}$ is on the left-hand side of the equation, estimation error will not lead to an attenuation of β.

Third, to understand the degree of variance inflation in the VA estimates, we report the ratio of the standard deviations of the estimated and true principal effects: $\sigma_{\tilde{\delta_j}}/\sigma_{\delta_J}$. A ratio of one would indicate that the distribution of principal VA estimates provides a good approximation of the magnitude of principal effects on student outcomes, while a ratio greater than (less than) one would indicate that the model overstates (understates) the magnitude of principal effects.

As discussed previously, principal VA models with school FE produce estimates of principal effects that are *local* to the connected network. It is nonetheless informative to also evaluate these estimates as *global* measures of principal effectiveness. While conceptually invalid, prior work often finds that VA estimators that ignore certain structural concerns can still perform well—or even better than estimators that are conceptually valid. Thus, we also compute these three summary measures for the *unadjusted* true principal effects (i.e., not accounting for connected networks).

Data and Network Structure

We apply the DGP described by (7) to simulated data created by a job separation and market clearing process, as well as actual student-level administrative data from Tennessee and NYC. For the simulated data, we specify a yearly turnover rate of 20%,

which corresponds to the national average. The basic simulation process is to (1) identify job separations, (2) determine whether the departing principal transfers to another opening or leaves the dataset, and (3) fill all openings with transferring principals and new-to-dataset principals. We repeat this process for 10 years across 1000 schools. We construct different simulated datasets by varying parameters that determine job separations, sorting to schools, and attrition from the dataset. However, we focus our main analysis on datasets with high versus low attrition, as this is the key parameter that affects the structure of connected networks. We provide a full description of the simulated labor markets and results in Appendix C.

For the empirical datasets, we can identify the linkages between students, schools, and principals, which allows us to test VA models using the actual connected networks of principals. For Tennessee, the analysis years run from 2007–2019 and include 3,835 principals, 1,719 schools, and roughly 5.1 million student-by-year observations. The NYC data goes from 1999–2017, containing 3,147 principals, 1,332 schools, and roughly 7.3 million student-by-year observations.

Table 2 summarizes the connected networks of principals in Tennessee and NYC, respectively. In Tennessee, there are 762 individual networks, though most either consist of a single school (74%) or are small (22.8%) which we define as a network with between two and five schools. Single-school networks have 2.5 principals while small networks have, 5.8 principals and 2.6 schools, on average. Tennessee also has 22 medium-sized networks (6–15 schools), comprising less than 3% of all individual networks, and two large networks (16+ schools). Six percent of principals have no network, meaning that they were the sole principal observed in a school across the analysis years.

By contrast, NYC has fewer principals in medium- or large-sized networks despite our access to a longer panel, and 63% of principals are in a single-school network. The average number of principals and schools in the single and small-sized networks in NYC are very similar to those found in Tennessee, as are the proportions of principals who have no

network. NYC's lack of principals in medium- and large-sized networks reflects an important difference in the principal labor markets between Tennessee and NYC; while we do observe some principals who transfer schools in Tennessee (which creates the necessary linkages for larger networks), the principal transfer rate in NYC is nearly zero. In other words, principals who leave their positions in NYC almost exclusively transfer out of the district (movement which we cannot observe with these data) or move out of the principalship entirely (e.g., retirements).

The final three rows of Table 2 concern the connectivity structure of the networks, which determines the precision of the FE estimates from the P+S model. Specifically, principals' VA estimates become more reliable as the network grows more dense, both in terms of the number of direct comparisons between different principals as well as the number of observations (i.e., student test scores). Following the approach of Jochmans and Weidner (2019), we analyze each network as a bipartite graph to produce the predicted amount of variance inflation (reported as a percentage of the error variance) based on its normalized Laplacian matrix. We also report the smallest nonzero eigenvalue (λ_2)—a global measure of connectivity where $\lambda_2 \to 0$ as the network becomes more sparse. Across all network sizes in both contexts, principal VA models benefit from large sample sizes, since all of the tested students in a school will contribute to estimating the principal's effect. Larger networks tend to be weakly connected, however, which will lead to non-trivial variance inflation in the VA estimates. In Tennessee's two large networks, for instance, the mean predicted variance inflation is 0.027, which is scaled in terms of the error variance of student test score growth.

Overall, Table 2 shows that the precision of VA estimates is very high for principals in small or medium-sized networks in Tennessee and NYC. Of course, the trade-off is that these VA estimates are highly localized measures of performance—the majority of principals can only be compared to other principals who worked in the same school. While there are a few large networks that, in theory, produce a more globalized measure of

performance, the underlying network structure is extremely weak, leading to VA estimates that contain considerable estimation error.

Results

Mobility From Simulated Labor Markets

We begin by showing results from our datasets with simulated mobility, where we can indirectly manipulate network size and connectivity. This provides a baseline to which we can compare with the actual networks formed by principals in Tennessee and NYC. A detailed explanation of these datasets is available in Appendix C, but the key parameter is the attrition rate, which determines whether principals leaving their schools exit the dataset entirely or move to another school.[15] Figure 2 shows results for our main summary measure—the correlation between the estimated and true principal effects—for a zero attrition (no principals leave) and high attrition (60% of turnover events are exits) labor market. In each case, the turnover rate is set at 20%, which is approximately equal the rate observed nationally. The high attrition scenario is representative of the typical principal labor market, while the low attrition scenario is ideal for generating sufficient mobility to yield strong connected networks.

In each panel, the performance measure for each model specification (defined by the legend at the bottom) is shown by the size of the principal–school match (top x-axis) and the correlation between the fixed principal and school effects (bottom x-axis). The left column shows correlations where the true principal effect is residualized within connected networks to correspond the "local" nature of the VA estimates. We also compare the VA estimates to the unadjusted principal effects (global) in the right column. This helps us understand the extent to which treating local measures of performance as global measures leads to inaccurate inferences about principal effectiveness.

[15] While we also examined datasets with endogenous separations and geographic localization of labor markets, these parameters were far less important for the performance of principal VA models (see Appendix C).

Figure 2 demonstrates that when attrition is low, whereby principals switch schools rather than exit the dataset, all of the estimators perform well in terms of their correlation with the true principal effects. In panel A, for instance, correlations in the no match scenario are all roughly 0.9 or above, except for a model that omits school effects. In this case, model performance is dependent on the correlation between the principal effect and (omitted) school effect. This scenario is ideal for estimating both principal and school effects because of the large number of switches that create strong connections among schools. Here, there is a single network that is well connected, which allows for global comparisons of principals with low variance inflation.[16]

The large match scenario of panel A shows that even in single a well-connected network, the presence of principal–school complementarities degrades the accuracy of network-based principal VA models. By contrast, match effects have negligible (P FE) or no (P-S+S FE) influence on the other models. This inaccuracy is a product of additional noise in the principal effect estimates rather than a systematic bias. Because the P+S FE model relies on mobile principals to identify the school effects, the addition of a principal-school match effect makes this source of variation inherently unreliable. Put another way, using the performance of principal A in two different schools to make indirect comparisons among other principals is problematic if principal A's true performance varies in these two schools, particularly if principal A is the *only* connection between the schools. In short, principal–school complementarities amplify the existing weakness of the connected networks approach for estimating principal effects. When match effects are present, estimators that explicitly include a principal-by-school effect (P-S+S FE or P+S+M RE) perform better in terms of measuring *localized* performance. These additional parameters come with the cost of restricting comparisons to principals who worked in the same school, thus creating a divergence between correlations in panel A (local comparisons)

[16] Because there is one network that includes virtually all of the principals, the P+S FE results are nearly identical in Panels A (local) and B (global).

and panel B (global comparisons). The P FE model will simply provide an estimate of the weighted average of principal effectiveness (i.e., the portable component and match effect) across the schools in which a principal worked.

Panels C and D of Figure 2 illustrate how high principal attrition rates create problems for estimating principal VA. Beginning with the P+S FE model, we observe in panel C that the correlation with the true principal effects is roughly 0.7 when no match effects exist—substantially lower than under the low attrition scenario. With match effects, this correlation drops to 0.5. The consequences of high attrition rates are that connected networks tend to be fragmented because of limited mobility, creating a combination of small networks (i.e., those with only a one or a few schools) and large networks that are weakly connected. Even when we focus on within-network performance (panel C), the estimates are substantially less reliable than in the low attrition scenario. High attrition also exacerbates the challenges of match effects because a larger portion of schools are connected through just one or two principals, whereby the school and match effects cannot be disentangled. Accordingly, the decrease in model performance for P+S FE between no match and large match scenarios is larger with high attrition rates than low attrition rates.

Given that the inclusion of school FE in principal VA models creates challenges—either in the form of noisy estimates or limited comparison sets—a clear question is whether to omit school FE entirely. The results in Figure 2 demonstrate, however, that if there is a fixed school contribution to test score growth, the exclusion of school FE (P FE model) can lead to wildly inaccurate VA estimates. This model is particularly inaccurate when the true principal and school effects are negatively correlated.

The sensitivity of the estimates to the inclusion of school FE is due to the fact that most principals are observed in only one or two schools. Any omitted school effect will be incorrectly attributed to the principal, producing an inaccurate estimate even if there is no absolute sorting of principals to schools (i.e., where the correlation between the fixed principal and school effect is zero). In essence, the issue is one of omitted variables bias.

The connected networks model is, effectively, a model with two large sets of indicator variables that are highly correlated with one another. Omitting the school indicators creates bias in each of the principal indicators, which grows in size with the magnitude of school effects. This bias exists in both low and high attrition scenarios, though it is greater under high attrition because the typical principal is only observed in a single school before exiting.

Importantly, models aimed at addressing the estimation error from weakly connected networks still perform quite well in the high attrition scenario. In panel C, for instance, we find that P-S+S FE and random effects shrinkage approaches under high attrition perform equal to or slightly worse than under low attrition. P-S+S FE, of course, avoids the attrition problem by not allowing for comparisons of principal across schools, meaning that it performs well for local comparisons but worse for global comparisons. Interestingly, however, it outperforms P+S FE even for global rankings of principals, demonstrating just how unreliable P+S FE becomes with weak networks. Explicitly modeling the match effects (P+S+M RE) produces estimates that are very similar to the P-S+S FE model in both local and global comparisons. Comparing P+S FE to its random effects equivalent shows how the shrinkage approach greatly improves the performance of principal VA models that leverage across-school comparisons of principals. In both panels C and D, P+S RE yields substantially stronger correlations between the estimated and true principal effects for both local and global comparisons.

Figures 3 and 4 show equivalent sets of results for the standard deviation ratio and bias measure. The former measures the extent to which the distribution of principal VA estimates provides an accurate indication of the true magnitude of principal effects. Here, it is important to note that if one's goal is purely to estimate the magnitude of principal effects—as opposed to producing individual-level estimates of effects—the model-based variance estimate from a random effects model is a more natural method (as opposed to examining the standard deviation of the distribution of estimates). Nonetheless, examining

the SD ratio is informative because prior studies in this literature tend to rely on the SD of the FE estimates to infer magnitude. Additionally, the SD ratio provides insight about whether low correlations between estimated and true effects are driven by variance inflation as opposed to bias.

Panel A of Figure 3 shows that under ideal circumstances, the SD ratio is consistently close to 1 across estimators, meaning that the empirical distribution of VA estimators provides a good estimate of the true distribution of principal effects. One caveat is that estimators that restrict the size of networks provide accurate indications of *within-network* magnitude, but will tend to understate the true variability of principal VA across the entire dataset. This is shown in panel B, where the P-S+S FE and P+S+M RE models yield a SD ratio below one. Consistent with the results in Figure 2, the presence of a principal–school match component leads to variance inflation in the VA estimates from the P+S FE model.

The key result previously shown in Figure 2 is that high attrition rates lead to weakly connected networks and, ultimately, inaccurate estimates of principals' effects from the P+S FE model. Together, Figures 3 and 4 help elucidate the source of this inaccuracy. Figure 4 shows the results of regressing the principal VA estimates on the true principal effects. In models with school FE, the estimated coefficient is roughly 0.9, with a small amount of bias introduced by controlling for the prior-year test score in models with principal and school effects. The prior-year test score also serves as a strong proxy for unobserved student heterogeneity, such that the small amount of bias is outweighed by the gain in precision from lower mean squared error.[17] Notably, the bias measure for the FE models is very similar in low attrition and high attrition scenarios. By comparison, Figure 3 shows that the P+S FE model contains substantially greater estimation error in the high

[17] For instance, omitting the prior-year test score results in substantially *worse* estimates in terms of the correlation with the true effects because of increased noise. We also estimated models where we constrain the coefficient on the lagged test score to its true value of 0.5 (as opposed to 0.7 in the actual results) and found effectively no change in the accuracy of the VA estimates.

attrition scenario. This noise component contains no information about principal effectiveness and thus lowers the accuracy of the P+S FE model.

Figures 3 and 4 also show the bias/variance tradeoff inherent to the random effects shrinkage approach. Specifically, the RE models yield a bias measure that is further from 1 but a SD ratio that is close to 1 regardless of whether network connectivity is weak or strong. Comparing panels A and C of Figure 4 is particularly informative in showing that with high attrition there is a greater need for shrinkage, which leads to a smaller bias measure (i.e., further from 1) in order to maintain an SD ratio of 1.

Overall, the results in Figure 2 using the simulated labor markets demonstrate that without a well-connected network of schools created by many principals switching schools, the connected networks model with principal and school FE that is widely used in prior work will produce inaccurate estimates of principal effects and will overstate the magnitude of principals' effects. This inaccuracy can be remedied by restricting the size of networks or by employing a random effects shrinkage estimator. Next, we examine how these principal VA simulations play out using the actual connected networks in Tennessee and New York City—both of which are labor markets with high principal attrition.

Results From Observed Labor Markets

Figure 5 shows results for each summary measure across simulations in Tennessee. Given the lack of variation in network sizes in NYC and similar results across contexts, we focus on the TN results with the NYC results shown in Appendix Figure A1.[18] The top row shows local performance (the true principal effects are residualized on network FE) and the bottom row shows global performance (no adjustments).

As demonstrated by Table 2, high rates of principal attrition in TN lead to fragmented and weakly connected networks of principals and schools, which should lead to

[18] The only substantive difference between TN and NYC is that the difference between the P+S FE and P-S+S FE models is smaller in NYC. Because the vast majority principals are already in small networks, limiting to single-school networks has less of an impact on the results relative to TN.

inaccuracy in the connected networks approach with principal and school FE. This is borne out in Figure 5, where the P+S FE approach yields low correlations with the true principal effects relative to the other estimators. In general, the results are remarkably similar to those from the high attrition simulated labor market in panels C and D of Figures 2–4. Given the weak network structure, the P-S+S FE or random effects models will yield better estimates of individual principals' effects and a more accurate indication of the magnitude of principal effects.

To reinforce the importance of network structure and the value of a shrinkage approach, we next compute our performance measures for the P+S FE (filled markers) and RE (hollow markers) models according to network size: single (1 school), small (2–5 schools), medium (6–15 schools), and large (16+ schools). Consistent with the network structure analysis in Table 2, Figure 6 shows that the estimation error in the P+S FE model is driven by larger networks. Specifically, Panel A shows that the correlation between estimated and true principal effects from large networks are much lower than those from smaller networks. In the no match scenario, correlations from small networks are roughly 0.9, while those from large networks are roughly 0.6.

Panels B and C show that this difference in accuracy between larger and smaller networks is completely driven by variance inflation. While there are no substantive differences by network size for P+S FE in the bias measure in panel B, the ratio of standard deviations in panel D are substantially greater for large networks. In the no match scenario in panel C, for example, the estimated SD of principal VA for single or small networks is approximately equal to the true within-network SD of principal effects, while the estimated SD for large networks is roughly 1.6 times larger than the true SD.

The results in Figure 6 demonstrate both a key tradeoff in the connected networks approach and a potential resolution. As the size of the connected networks grows, principal VA estimates become less localized, which increases their usefulness as a measure of principal effectiveness. At the same time, the reliability of the VA estimates decreases due

to weak network structure. Thus, the ability to link an increasing number of principals is not unambiguously beneficial even if the underlying assumptions of the principal and school fixed effects model are met. This can be remedied, however, by alleviating through shrinkage the estimation in error in large networks. Parallel to the results in the aggregate, the P+S RE trades off bias for lower variance inflation, which ultimately increases the correlation between principal VA estimates and the true principal effects. Figure 6 reinforces that the value of the shrinkage approach is concentrated in large networks, whereas there is little change in the accuracy of estimates between FE and RE in small networks.

Empirical Application

As a final supplement to our simulation results, we provide a brief demonstration using principal VA estimates from actual student test scores. Specifically, we estimate the P+S FE, P-S+S FE, and P+S RE models, with the distribution of VA estimates summarized in Table 3.[19] In addition to the full sample, we show the SD of the estimates for principals by network size. We again focus on the TN results, with the NYC results shown in Appendix Table A1.

As our simulation results would suggest, the SD of the P+S FE model estimates for actual math and reading VA increases substantially with network size, reflecting the variance inflation due to sparse network structure. For instance, the SD of math VA in single-school networks is 0.09, meaning that a 1 SD increase in principal VA raises math test scores by 0.09 student-level SD, on average. In large networks, this SD increases to 0.32. Turning to the random effects and P-S+S FE models, which should contain little to no estimation error, we observe that the estimated magnitude of principal VA does increase

[19] In the control vector, we include a broader set of student and school characteristics to align with common principal VA specifications. Specifically, we control for cubics of prior-year test scores in math and reading, prior-year attendance rate, student demographic characteristics (race/ethnicity, gender, economically disadvantaged, English learner, gifted classification, special education classification, and grade repetition), and school-by-year averages of these demographics.

in larger networks, but by substantially less than would be implied by the P+S FE results. The mixed model—which in our simulations provides the most accurate estimates of the magnitude of principal VA—suggests that the true SD in large networks is 0.14 for math and 0.06 for reading. As previously noted in our simulation results, the within-network SD of principal VA in small networks will understate the true SD of principal quality, as the school FE sweeps out real differences in average principal quality between schools.

In an effort to examine the accuracy of principal VA across these specifications, we compare VA to plausible alternative measures of principal performance. We do this in two ways. First, we predict residualized student test scores in one subject as a function of principal VA in the other subject (e.g., predicting math scores using a principal's reading VA). Second, we compare VA estimates to principals' ratings from their supervisors.[20] Appendix Figures A2 and A3 show binned scatterplots for predicting the given performance measure as a function of their math or reading VA estimate across model specifications. These plots are consistent with the simulation results in showing how large estimation error attenuates the correlation between P+S FE VA estimates and the performance measure. While there is an apparent upward slope when restricting to principals in the middle of the distribution, the high-leverage observations at the tails (i.e., principals with very high or low VA estimates) flatten the estimated regression line. This estimation error can be reduced through a shrinkage approach or by limiting the size of connected networks. In both approaches, the distribution of VA is compressed and the estimated correlations with the performance measure increase.

[20] In Tennessee, principals beginning in 2011–12 receive rubric-based ratings from their supervisors as part of the state's high-stakes educator accountability system, where the average score comprises 50% of a principal's summative evaluation rating. Prior work has documented positive, though weak, relationships between supervisor ratings and principal VA (Bartanen, 2020; Grissom, Blissett, & Mitani, 2018). We hypothesize, however, that these correlations may be somewhat attenuated due to estimation error in the principal VA estimates. While principals receive an average observation score each year, we construct a time-invariant measure that averages across all available years. When comparing to the principal-by-school VA estimates, we limit to observation scores that were received in that same school.

Conclusion

There is a growing interest in using value-added models to estimate principals' contributions to student outcomes. Accurately isolating principal effectiveness requires accounting for school-level factors that affect student achievement but are not controlled by principals. The common approach to address this issue is to estimate a "connected networks" model with principal and school fixed effects. The accuracy of this model, however, has not been rigorously tested. Specifically, the inclusion of school FE—while conceptually important for mitigating bias—creates challenges with respect to the reliability of principal VA estimates.

Using simulated test scores applied to the actual principal–school assignments across long panels from Tennessee and New York City, we reach several important findings. First, limited mobility of principals combined with high rates of attrition makes the connected networks model difficult to implement. There is insufficient variation to jointly identify both principal and school effects, such that principal VA estimates are either highly localized—reflecting performance relative to only a handful of other principals—or very imprecise. In both Tennessee and New York City, the modal principal is in a single-school connected network. While estimates from small networks are reliable, they are less useful from a practical perspective and they understate the magnitude of principals' effects. On the other hand, VA estimates from large networks reflect a principal's performance relative to a much larger group, but the underlying network structure is weak. As a result, VA estimates from large networks are unreliable and their variance overstates the magnitude of principals' effects. The precision of VA estimates in large connected networks can be improved by employing a shrinkage estimator, though our simulation results suggest that a mixed model performs substantially better than a post-hoc shrinkage of the principal fixed effects, the latter of which is more common in the extant literature.

Our results help to inform the estimation of principal VA. Those implementing principal VA models should consider the intended use of the estimates when choosing the

most appropriate specification. For most applications, random effects models are the best available option. This is particularly true when the magnitude of principal effects is the main parameter of interest. While, models with principal-by-school and school FE produce VA estimates with stronger internal validity and higher reliability, they understate the importance of principal quality and the measures are highly localized. The inability to compare principals across schools, in particular, makes this model unfavorable for accountability purposes. A final alternative is to omit school FE entirely from the model. While this alleviates variance inflation and small network issues, it comes with a risk of substantial bias in the VA estimates, as any unobserved school-level factors that are not completely captured by observable school characteristics will be mistakenly attributed to principals.

Given our results, a clear question is whether principal VA can really provide useful information about principal effectiveness. Certainly, the multi-faceted and indirect nature of principals' contributions to student outcomes makes estimating principal effects a formidable challenge. We stress, however, that even imperfect principal VA models may contain valuable information that is not captured by alternative measures. A clear strength of the principal effects literature, for instance, is the ability to avoid penalizing principals who work in the most challenging schools. This is particularly important given evidence that rubric-based ratings of principal practice and school value-added—two commonly used alternative measures of principal performance—in part hold principals accountable for factors they cannot control (Chiang et al., 2016; Grissom et al., 2018, 2015).

As a final caution, we note that our simulation analysis makes a number of assumptions about the nature of principal effects on student outcomes. These assumptions—most notably, that new principals can immediately change school performance and their effects are fixed over time—are helpful for isolating issues related to the connected networks approach, but may not hold in practice. While this study is an additional step in understanding the extent to which principal VA models can provide

accurate estimates of principals' effects, there is a continued need for work that rigorously examines their validity and reliability. Future work should continue to outline the various assumptions behind principal VA models and test these assumptions in both simulation and empirical analyses.

Table 1
Simulation Parameters

Panel A: Data-Generating Process

$A_{ijst} = \lambda A_{ijs,t-1} + \theta_{jst} + c_i + e_{ijst}$, where
$\theta_{jst} = \delta_{j(t,s)} + \alpha_{js(t,s)} + \gamma_s + v_{jst}$

Panel B: Parameters that are Fixed Across Simulations

School Performance Effect (θ_{jst}) % of Total Variance	5%
λ (persistence)	0.5
$A_{ijs,t-n}$ (base score)	$Normal(0, 1)$
c_i (fixed student effect)	$Normal(0, 0.5)$
e_{ijst} (random deviation)	$Normal(0, 1)$
Corr$(A_{ijs,t-n}, c_i)$	0.5

Panel C: Parameters that Vary Across Simulations

Component Shares of School Performance Effect				Corr(δ_j, γ_s)
δ_j	α_{js}	γ_s	v_{jst}	
0.45	0.00	0.45	0.10	-0.4
0.20	0.25	0.45	0.10	0
				0.4

Notes: A denotes the test score for student i with principal j and school s in year t. The base score, $A_{ijs,t-n}$, is randomly drawn for a student in their first observed year in the dataset because no prior-year score is observed. As denoted by Corr$(A_{ijs,t-n}, c_i)$, the base score is drawn to have a 0.5 correlation with the fixed student effect.

Table 2

Network Statistics for Tennessee and New York Principals

	Network Size (# of Schools)				
	Single (1)	Small (2–5)	Medium (6–15)	Large (16+)	No Network
Panel A: Tennessee					
Number of Networks	564	174	22	2	
Mean Schools per Network	1	2.6	8.3	142.5	
Mean Principals per Network	2.5	5.8	20.5	353	
Number of Principals	1435	1003	450	706	243
Percentage of Total Principals	37.4	26.2	11.7	18.4	6.3
Mean Student Obs per Prin	1267	1457	1379	1470	2069
Connectivity (λ_2)	1.6664	0.6148	0.0438	0.0014	
Variance Inflation	0.002	0.004	0.007	0.027	
Panel B: New York City					
Number of Networks	720	107	10	2	
Mean Schools per Network	1	2.5	8.1	20.5	
Mean Principals per Network	2.7	6.3	18.9	47	
Number of Principals	1970	671	189	94	224
Percentage of Total Principals	62.6	21.3	6.0	3.0	7.1
Mean Student Obs per Prin	2552	2321	2272	1930	2150
Connectivity (λ_2)	1.5715	0.5209	0.0414	0.0073	
Variance Inflation	0.002	0.003	0.005	0.010	

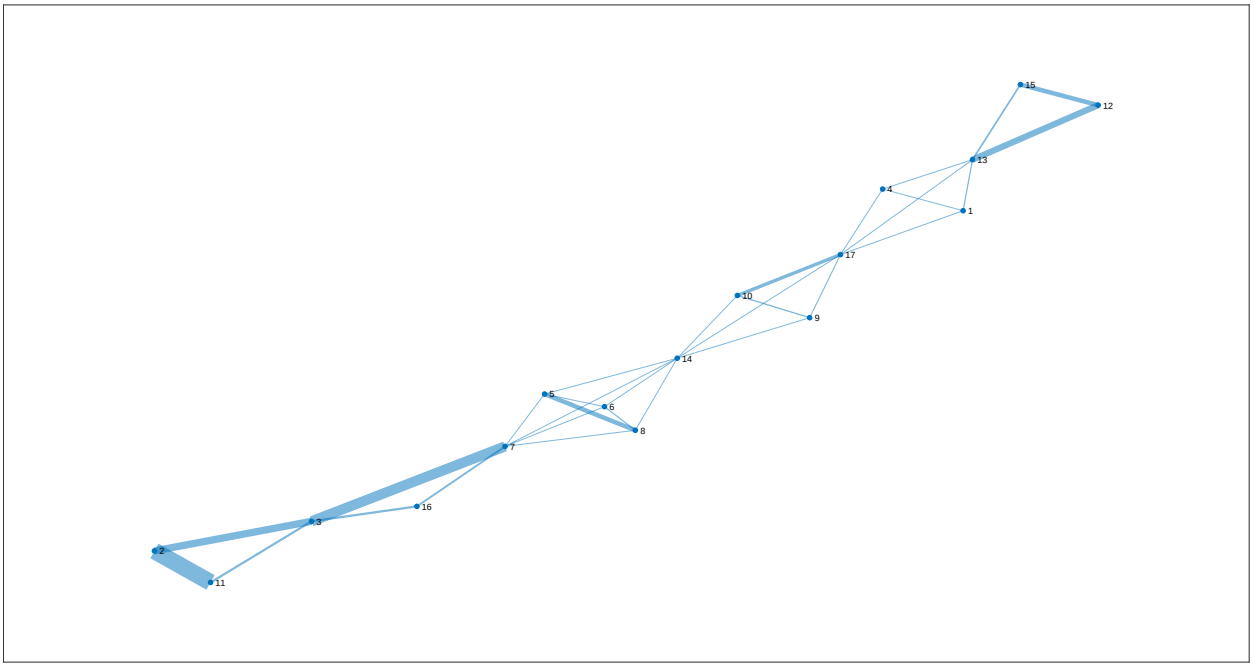
Notes: Both connectivity (λ_2) and variance inflation are calculated following the approach of Jochmans and Weidner (2019). λ_2 is the smallest non-zero eigenvalue from the normalized Laplacian matrix that corresponds to each connected graph of principals. A smaller eigenvalue indicates that principals in a network are more weakly connected. Variance inflation is expressed in terms of the error variance of student test score growth. Principals without a network are those who were the only principal in their school across the study period.

Table 3

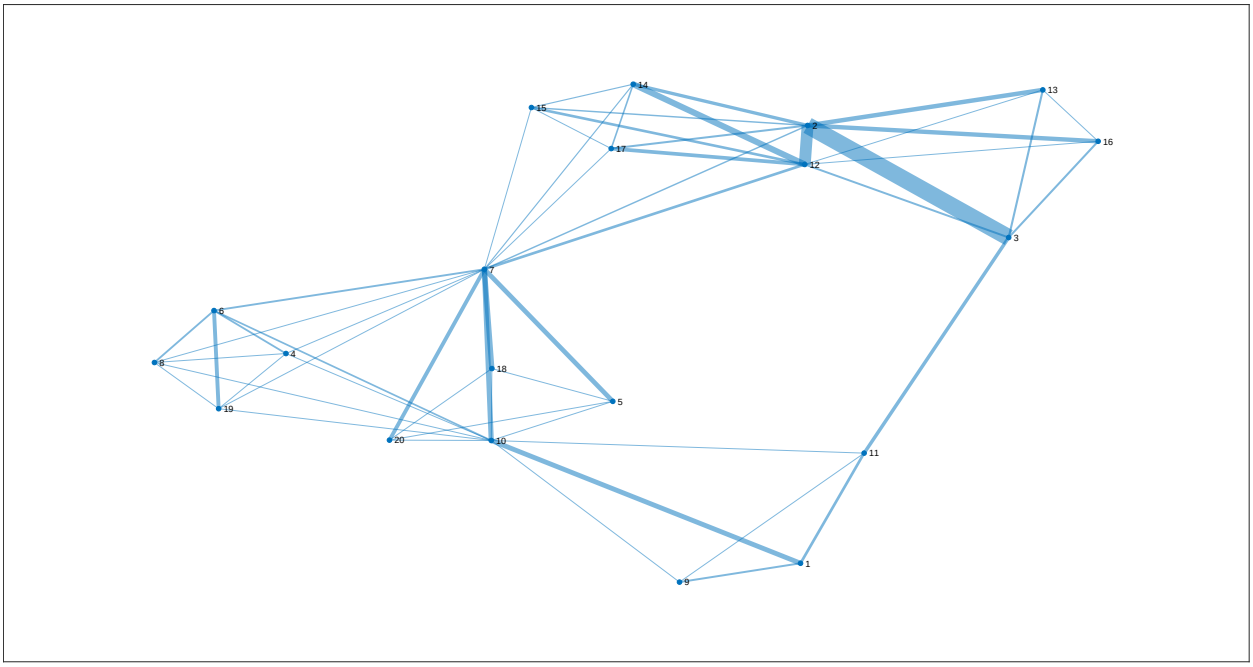
Standard Deviation of Empirical Principal VA Estimates in Tennessee

		Network Size (# of Schools)			
	All	Single (1)	Small (2–5)	Medium (6–15)	Large (16+)
Math					
Prin + School FE	0.201	0.094	0.152	0.259	0.323
Principal RE + School RE	0.129	0.091	0.126	0.160	0.166
Prin-School + School FE	0.099	0.084	0.101	0.110	0.114
Reading					
Prin + School FE	0.127	0.054	0.083	0.195	0.203
Principal RE + School RE	0.053	0.040	0.055	0.062	0.066
Prin-School + School FE	0.059	0.048	0.057	0.069	0.072

Notes: Table shows the standard deviation of principal VA estimates using actual student test scores. For the mixed model, the VA estimates are the best linear unbiased predictions (BLUPs) from a model with school fixed effects and principal random effects. The model-based estimate of the SD of the principal random effect is 0.148 in math and 0.066 in reading. The BLUPs are less variable due to shrinkage.



(a) *Weaker Connected Network ($\lambda_2 = 0.008$, Variance Inflation $= 0.022$)*



(b) *Stronger Connected Network ($\lambda_2 = 0.114$, Variance Inflation $= 0.005$)*

Figure 1

Examples of Connected Networks of Principals

Notes: Each plot shows a single connected network of principals from Tennessee. Nodes represent principals and edges are formed by principals who worked in the same school. In panel A, there are 20 principals across 6 schools. In panel B, there are 17 principals across 6 schools. The weight (shown visually by width) of the edge is determined by the harmonic mean of the number of students that contribute to estimating each principal's effect in the relevant school. Both connectivity (λ_2) and variance inflation are calculated following the approach of Jochmans and Weidner (2019). λ_2 is the smallest non-zero eigenvalue from the normalized Laplacian matrix that corresponds to each connected graph of principals. A smaller eigenvalue indicates that principals in a network are more weakly connected. Variance inflation is expressed in terms of the error variance of student test score growth.

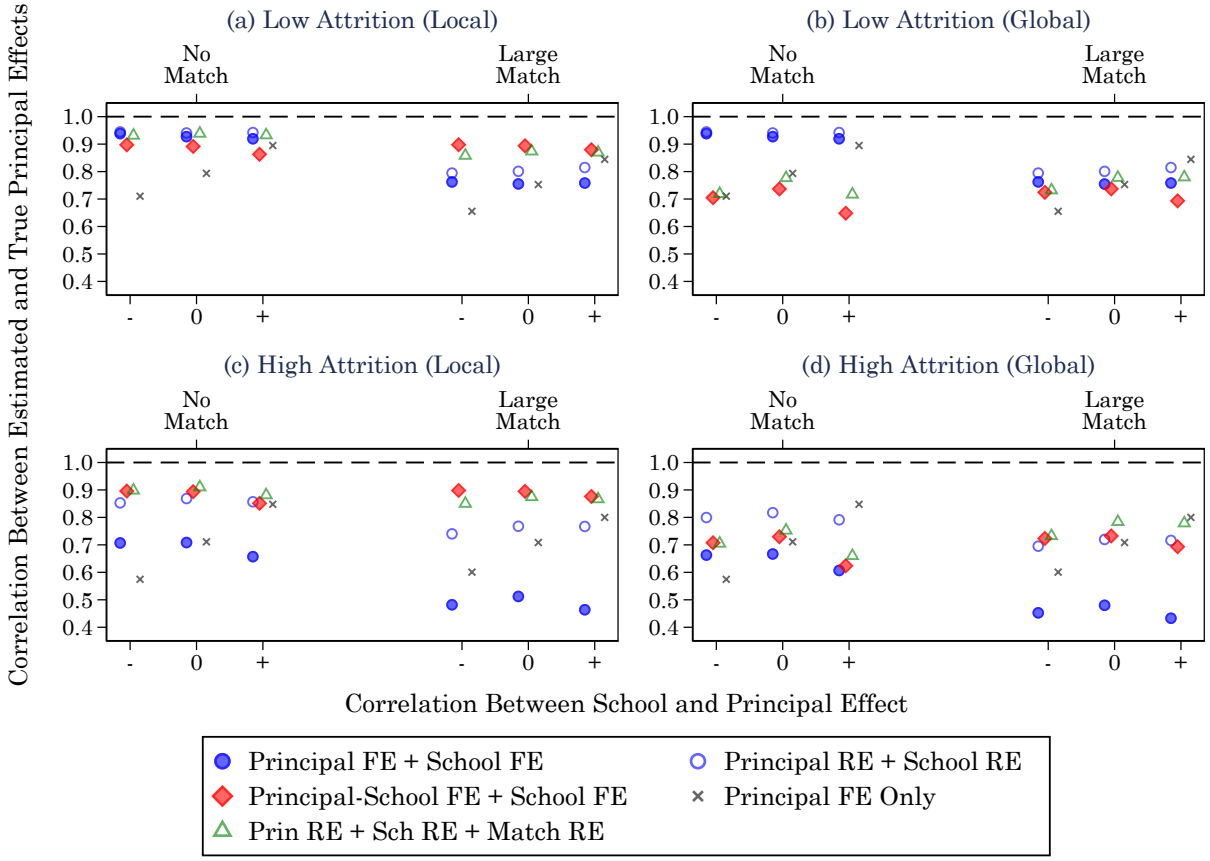


Figure 2

Results Using Simulated Labor Markets

Notes: The plot header denotes whether the simulated dataset is low attrition or high attrition. The bottom x-axis corresponds to the correlation between principal quality and the fixed school effect (-0.4, 0, 0.4). The top y-axis corresponds to the magnitude of the principal–school match effect (0% or 56% of the total principal effect). The school performance effect constitutes 5% of the total variance in student test score growth. In Panels A, B, and C, the true principal effects are residualized on network fixed effects corresponding to the connected networks formed by principals and schools for the given VA measure. For ease of interpretation, a small amount of horizontal spacing is added between models.

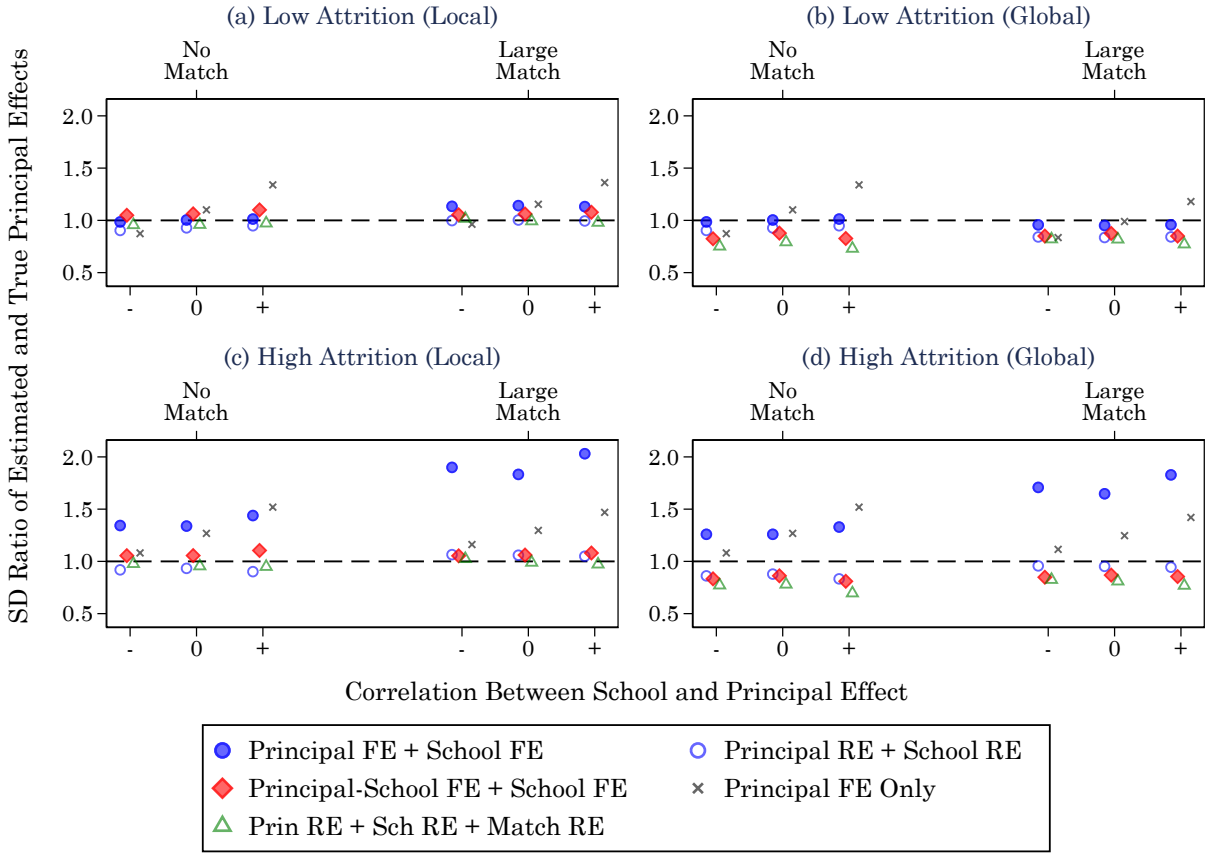


Figure 3

Results Using Simulated Labor Markets (SD Ratio)

Notes: The plot header denotes whether the simulated dataset is low attrition or high attrition. The bottom x-axis corresponds to the correlation between principal quality and the fixed school effect (-0.4, 0, 0.4). The top y-axis corresponds to the magnitude of the principal–school match effect (0% or 56% of the total principal effect). The school performance effect constitutes 5% of the total variance in student test score growth. In Panels A, B, and C, the true principal effects are residualized on network fixed effects corresponding to the connected networks formed by principals and schools for the given VA measure. For ease of interpretation, a small amount of horizontal spacing is added between models.

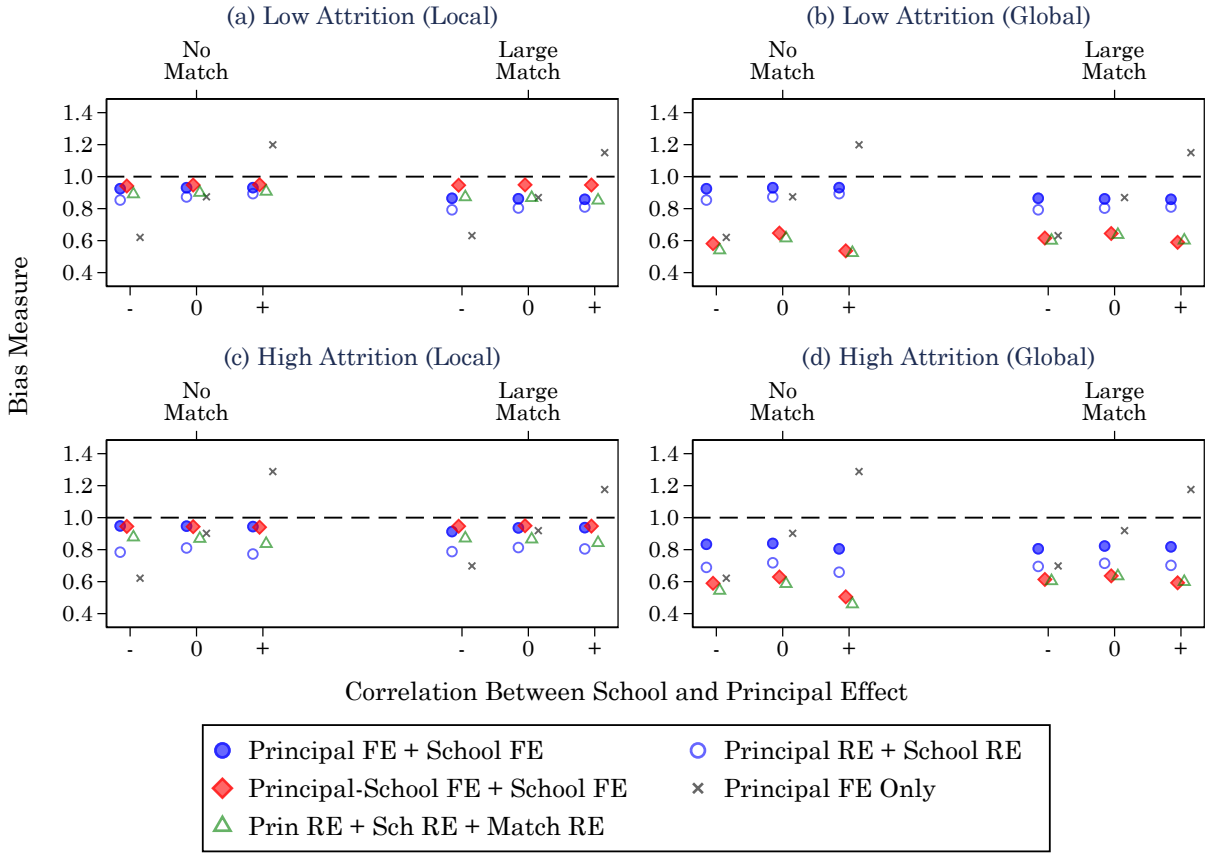


Figure 4

Results Using Simulated Labor Markets (Bias Measure)

Notes: The plot header denotes whether the simulated dataset is low attrition or high attrition. The bottom x-axis corresponds to the correlation between principal quality and the fixed school effect (-0.4, 0, 0.4). The top y-axis corresponds to the magnitude of the principal–school match effect (0% or 56% of the total principal effect). The school performance effect constitutes 5% of the total variance in student test score growth. In Panels A, B, and C, the true principal effects are residualized on network fixed effects corresponding to the connected networks formed by principals and schools for the given VA measure. For ease of interpretation, a small amount of horizontal spacing is added between models.

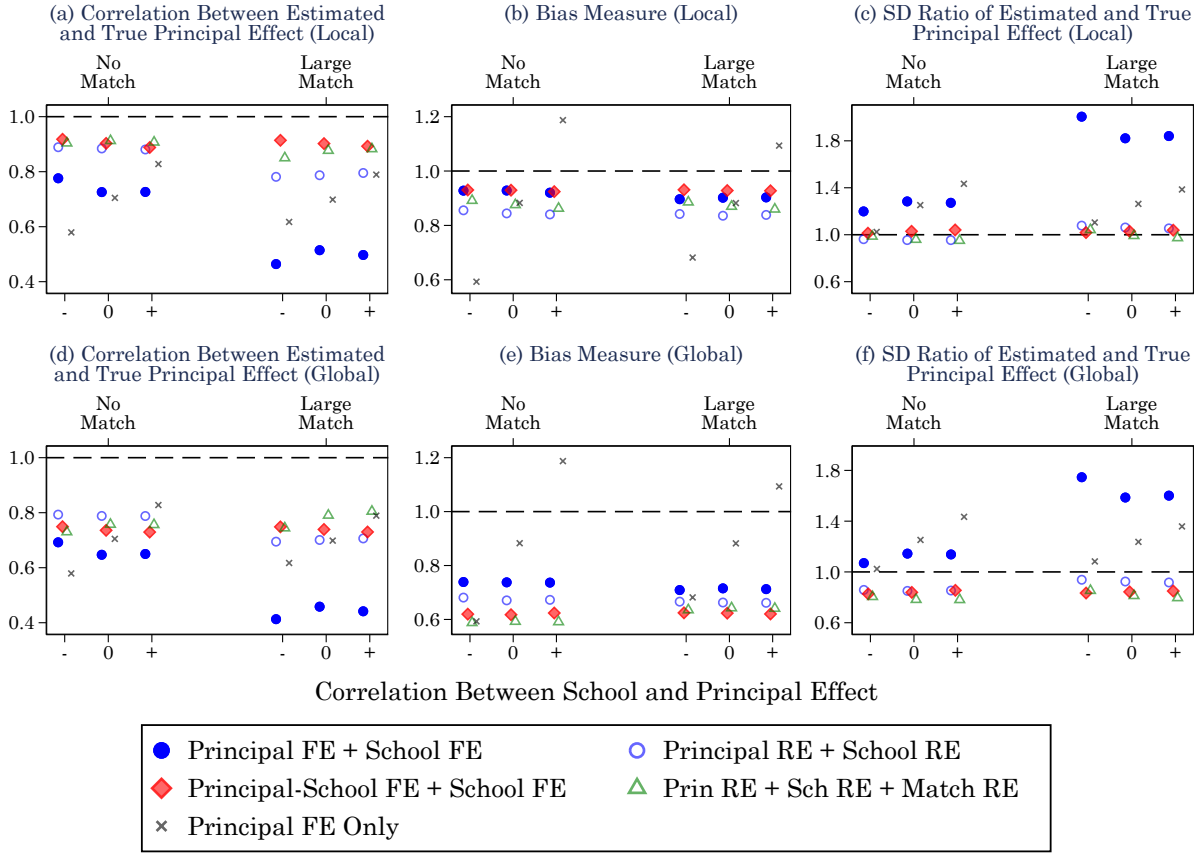


Figure 5

Results Using Tennessee Labor Markets

Notes: In each plot, the y-axis is defined by the header. The bottom x-axis corresponds to the correlation between principal quality and the fixed school effect (-0.4, 0, 0.4). The top y-axis corresponds to the magnitude of the principal–school match effect (0% or 56% of the total principal effect). The school performance effect constitutes 5% of the total variance in student test score growth. In Panels A, B, and C, the true principal effects are residualized on network fixed effects corresponding to the connected networks formed by principals and schools for the given VA measure. For ease of interpretation, a small amount of horizontal spacing is added between models.

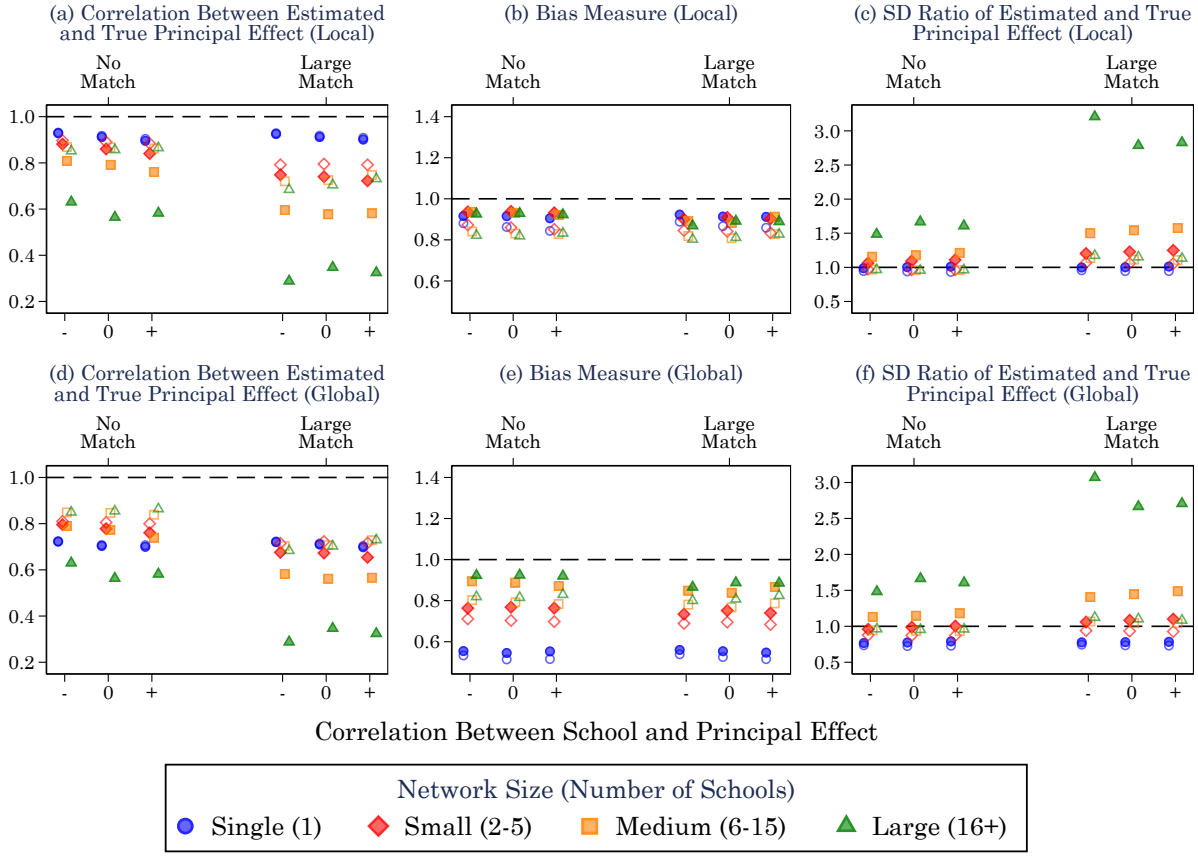


Figure 6

Principal + School FE and RE Results by Network Size

Notes: Results shown are only for the principal and school FE and RE models in Tennessee. Filled markers denote FE estimates and hollow markers denote RE estimates. The legend defines the size of the connected network. In each plot, the y-axis is defined by the header. The bottom x-axis corresponds to the correlation between principal quality and the fixed school effect (-0.4, 0, 0.4). The top y-axis corresponds to the magnitude of the principal–school match effect (0% or 56% of the total principal effect). In the results shown here, the school performance effect constitutes 5% of the total variance in student test score growth. In Panels A, B, and C, the true principal effects are residualized on network fixed effects corresponding to the connected networks formed by principals and schools for the given VA measure. For ease of interpretation, a small amount of horizontal spacing is added between models.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, *25*(1), 95–135.

Andrabi, T., Das, J., Khwaja, A. I., & Zajonc, T. (2011). Do value-added estimates add value? Accounting for learning dynamics. *American Economic Journal: Applied Economics*, *3*(3), 29–54.

Bartanen, B. (2020). Principal Quality and Student Attendance. *Educational Researcher*, *49*(2), 101–113.

Bartanen, B., & Grissom, J. A. (2021). School Principal Race, Teacher Racial Diversity, and Student Achievement. *Journal of Human Resources*.

Branch, G. F., Hanushek, E. A., & Rivkin, S. G. (2012). *Estimating the Effect of Leaders on Public Sector Productivity: The Case of School Principals.* Cambridge, MA.

Burkhauser, S. (2017). How Much Do School Principals Matter When It Comes to Teacher Working Conditions? *Educational Evaluation and Policy Analysis*, *39*(1), 126–145.

Chiang, H., Lipscomb, S., & Gill, B. (2016). Is School Value Added Indicative of Principal Quality? *Education Finance and Policy*, *11*(3), 283–309.

Dhuey, E., & Smith, J. (2018). How school principals influence student learning. *Empirical Economics*, *54*, 851–882.

Grissom, J. A., & Bartanen, B. (2019). Principal Effectiveness and Principal Turnover. *Education Finance and Policy*, *14*(3), 355–382.

Grissom, J. A., Bartanen, B., & Mitani, H. (2019). Principal Sorting and the Distribution of Principal Quality. *AERA Open*, *5*(2), 1–21.

Grissom, J. A., Blissett, R. S. L., & Mitani, H. (2018). Evaluating School Principals: Supervisor Ratings of Principal Practice and Principal Job Performance. *Educational Evaluation and Policy Analysis*, *40*(3), 446–472.

Grissom, J. A., Kalogrides, D., & Loeb, S. (2015). Using Student Test Scores to Measure Principal Performance. *Educational Evaluation and Policy Analysis*, *37*(1), 3–28.

Guarino, C. M., Maxfield, M., Reckase, M. D., Thompson, P. N., & Wooldridge, J. M. (2015). An Evaluation of Empirical Bayes's Estimation of Value-Added Teacher Performance Measures. *Journal of Educational and Behavioral Statistics*, *40*(2), 190–222.

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2015). Can Value-Added Measures of Teacher Performance Be Trusted? *Education Finance and Policy*, *10*(1), 117–156.

Hallinger, P., & Heck, R. H. (1998). Exploring the Principal's Contribution to School Effectiveness: 1980-1995. *School Effectiveness and School Improvement*, *9*(2), 157–191.

Husain, A. N., Matsa, D. A., & Miller, A. R. (2018). Do Male Workers Prefer Male Leaders? An Analysis of Principals' Effects on Teacher Retention. *NBER Working Paper Series*, 38.

Jackson, C. K. (2013). Match Quality, Worker Productivity, and Worker Mobility: Direct Evidence from Teachers. *The Review of Economics and Statistics*, *95*(4), 1096–1116.

Jochmans, K., & Weidner, M. (2019). Fixed-Effect Regressions on Network Data. *Econometrica*, *87*(5), 1543–1560.

Kline, P., Saggio, R., & Sølvsten, M. (2018). Leave-out estimation of variance components.

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review*, *47*, 180–195.

Mansfield, R. K. (2015). Teacher Quality and Student Inequality. *Journal of Labor Economics*, *33*(3), 751–788.

Mihaly, K., McCaffrey, D., Sass, T. R., & Lockwood, J. R. (2013). Where you come from or where you go? Distinguishing between school quality and the effectiveness of teacher preparation program graduates. *Education Finance and Policy*, *8*(4), 459–493.

Mihaly, K., McCaffrey, D. F., Lockwood, J., & Sass, T. R. (2010). Centering and reference groups for estimates of fixed effects: Modifications to felsdvreg. *The Stata Journal*, *10*(1), 82–103.

Sass, T. R., Semykina, A., & Harris, D. N. (2014). Value-added models and the measurement of teacher productivity. *Economics of Education Review*, *38*, 9–23.

Sebastian, J., & Allensworth, E. (2012). The Influence of Principal Leadership on Classroom Instruction and Student Learning: A Study of Mediated Pathways to Learning. *Educational Administration Quarterly*, *48*(4), 626–663.

Verdier, V. (2018). Estimation and Inference for Linear Models with Two-Way Fixed Effects and Sparsely Matched Data. *The Review of Economics and Statistics*, 1–38.

Woodcock, S. D. (2015). Match effects. *Research in Economics*, *69*(1), 100–121.

Appendix A
Appendix Figures and Tables

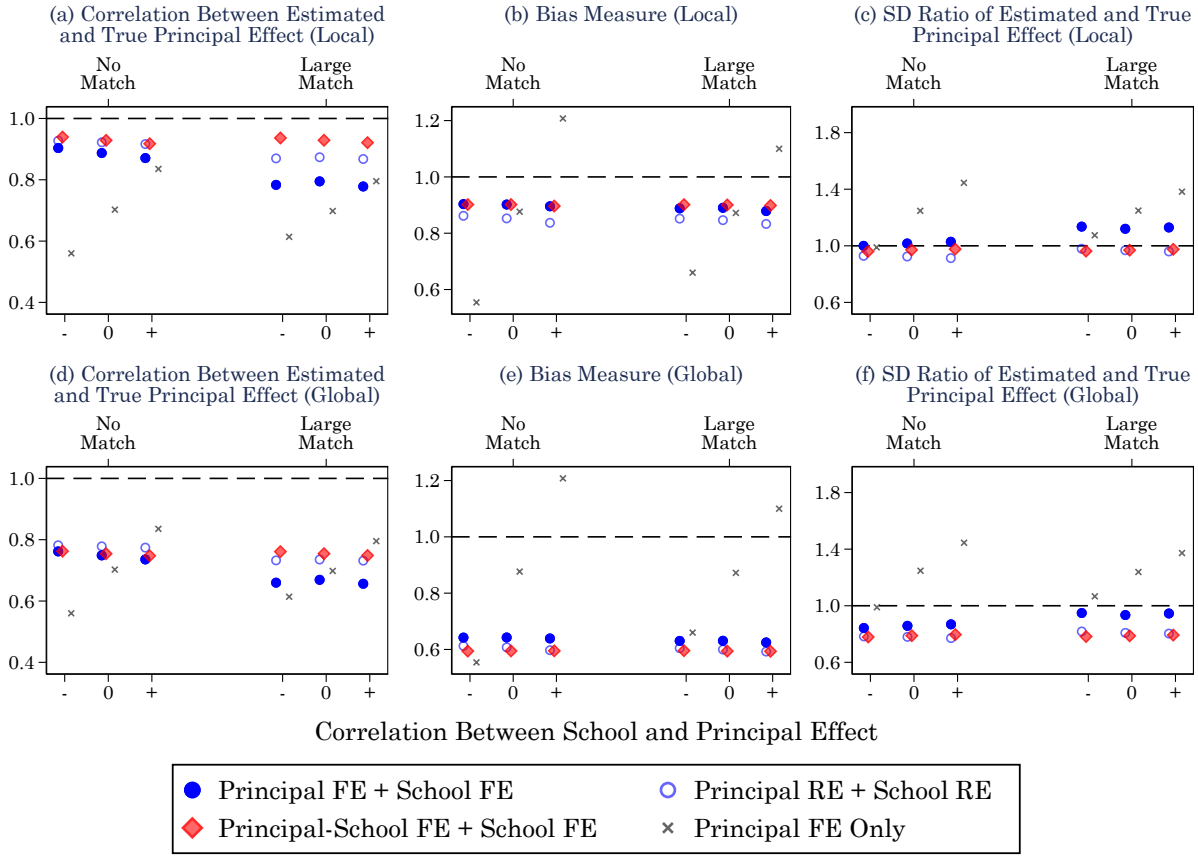


Figure A1

Results Using New York City Labor Market

Notes: In each plot, the y-axis is defined by the header. The bottom x-axis corresponds to the correlation between principal quality and the fixed school effect (-0.4, 0, 0.4). The top y-axis corresponds to the magnitude of the principal–school match effect (0% or 56% of the total principal effect). The school performance effect constitutes 5% of the total variance in student test score growth. In Panels A, B, and C, the true principal effects are residualized on network fixed effects corresponding to the connected networks formed by principals and schools for the given VA measure. For ease of interpretation, a small amount of horizontal spacing is added between models.

Table A1

Standard Deviation of Empirical Principal VA Estimates in New York City

	All	Network Size (# of Schools)			
		Single (1)	Small (2–5)	Medium (6–15)	Large (16+)
Math					
Principal FE + School FE	0.070	0.044	0.088	0.144	0.146
Principal RE + School RE	0.059	0.052	0.071	0.075	0.090
Principal-School FE + School FE	0.048	0.044	0.056	0.052	0.064
Reading					
Principal FE + School FE	0.051	0.033	0.065	0.098	0.143
Principal RE + School RE	0.040	0.037	0.046	0.043	0.047
Principal-School FE + School FE	0.035	0.033	0.039	0.037	0.041

Notes: Table shows the standard deviation of principal VA estimates using actual student test scores. For the mixed model, the VA estimates are the best linear unbiased predictions (BLUPs) from a model with school fixed effects and principal random effects. The model-based estimate of the SD of the principal random effect is 0.082 in math and 0.057 in reading. The BLUPs are less variable due to shrinkage.

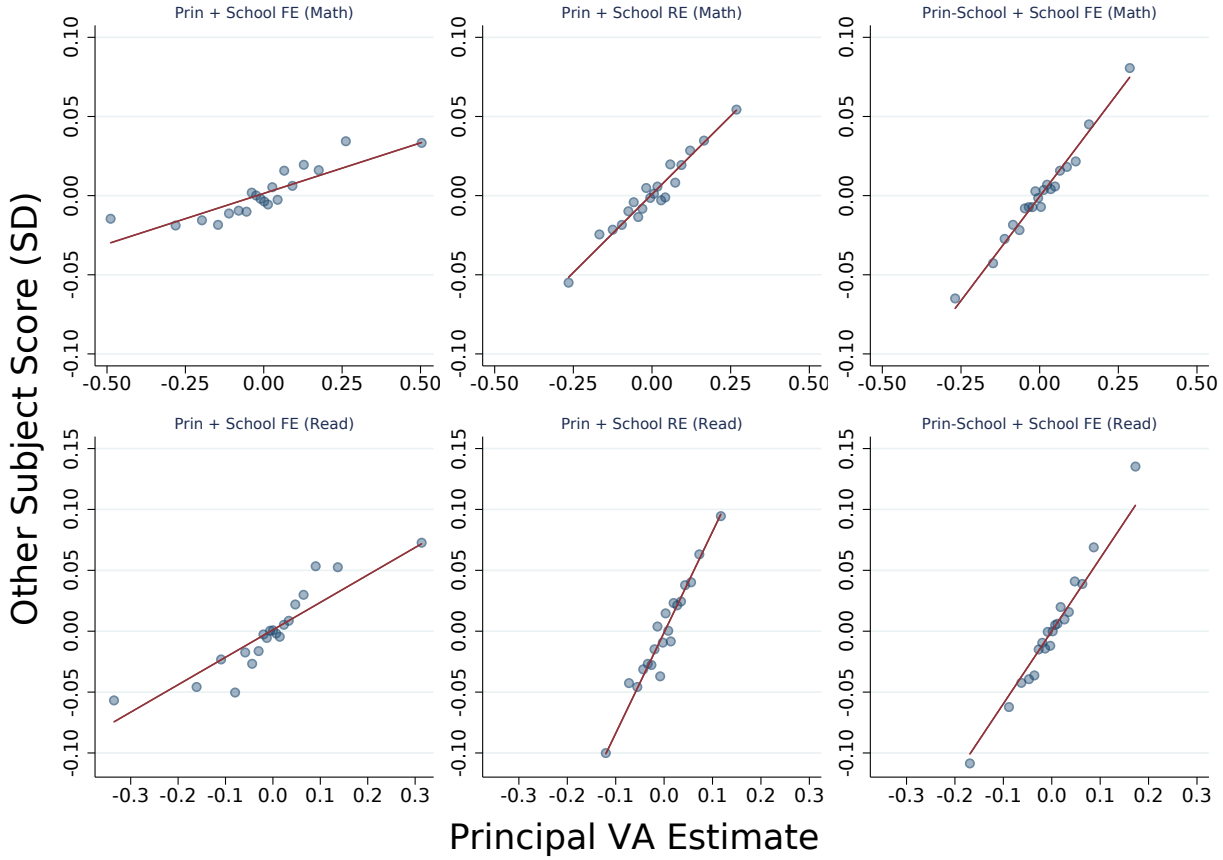


Figure A2

Relationship Between Principal VA Estimates and Opposite-Subject Test Scores

Notes: Each plot shows a binned scatterplot predicting student test scores in the opposite subject as a function of the principal VA estimate for the subject shown in the plot header, along with the OLS line. Models include controls for students' prior-year test scores, demographic characteristics, school-by-year averages of these characteristics, fixed effects for the connected network corresponding to the principal VA estimate.

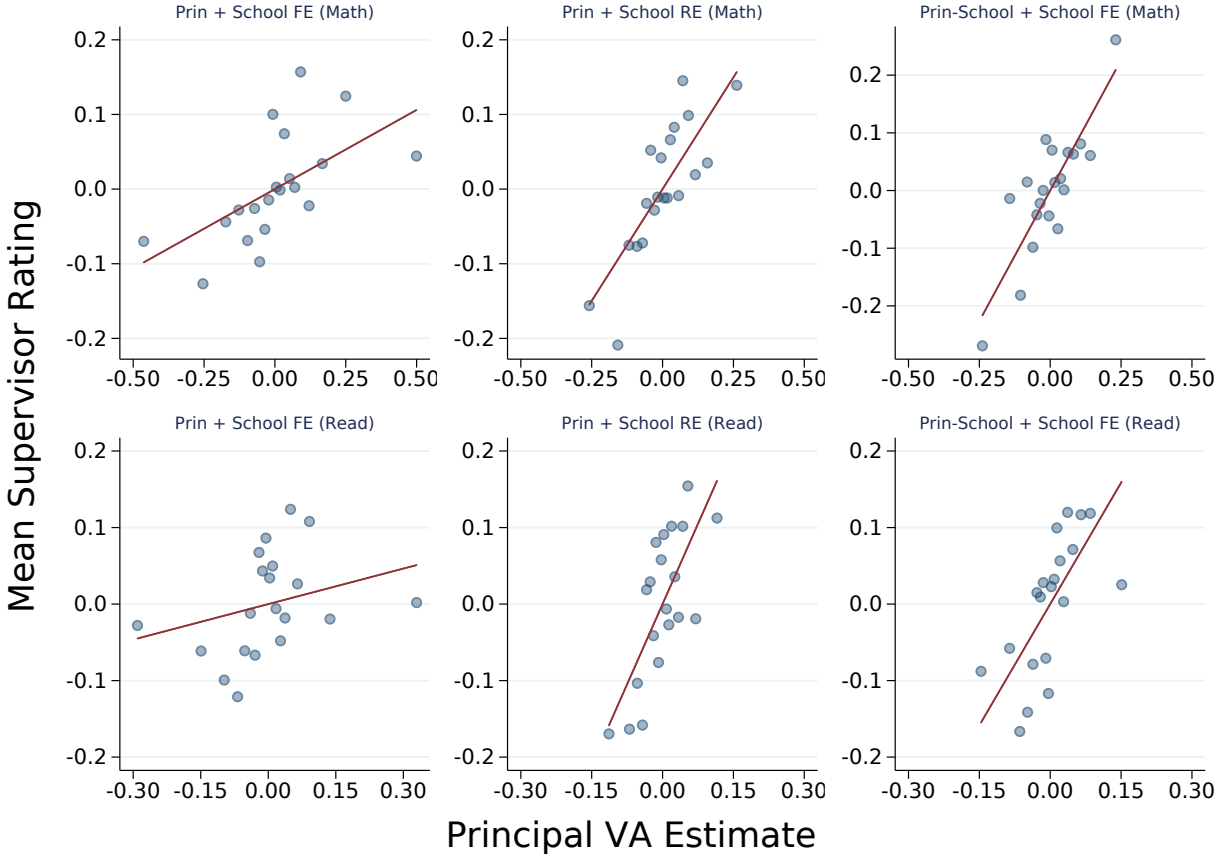


Figure A3

Relationship Between Principal VA Estimates and Supervisor Ratings in Tennessee

Notes: Each plot shows a binned scatterplot predicting supervisor ratings as a function of the principal VA estimate shown in the plot header, along with the OLS line. Models include fixed effects for the connected network corresponding to the principal VA estimate. The red line is estimated via OLS using the underlying data. Starting from the left, the standardized regression coefficient for the slope is (top row) 0.059, 0.105, 0.158; (bottom row) 0.031, 0.109, 0.116.

Appendix B
Variance Decomposition of Math Test Scores in Tennessee and New York City

To provide empirical support for our simulation parameters, we conducted variance decompositions for standardized math test scores in Tennessee and New York City. Specifically, we estimate via restricted maximum likelihood the following general form mixed model:

$$Y_{ijst} = \lambda(f(Y_{i,t-1})) + \gamma \mathbf{X}_{it} + \phi \mathbf{Z}_{st} + v_{ijst} \quad (1)$$

where v_{ijst} is a composite error term that includes either a (1) school-by-year random effect or (2) principal and school random effects. We control for prior-year test scores in both subjects and prior-year attendance rate (including squared and cubed terms), student characteristics (race/ethnicity, gender, FRPL-eligible, special education status, gifted status, flag for grade repetition, flag for within-year move to another school), and school-by-year means of the student characteristics.

Table B1
Variance Components for Student Test Score Growth

	Tennessee				New York City			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A								
School-by-Year	18.6	9.8	9.6	8.5	27.0	8.0	6.5	5.4
Residual	81.4	90.2	90.4	91.5	73.0	92.0	93.5	94.6
Panel B								
Principal	4.1	5.8	5.9	6.0	3.4	2.2	2.3	2.2
School	18.0	6.1	5.7	6.2	24.0	4.6	3.0	2.0
Residual	77.9	88.2	88.5	87.8	72.6	93.2	94.7	95.8
Prior-year Test Scores		✓	✓	✓		✓	✓	✓
Student Characteristics			✓	✓			✓	✓
School Characteristics				✓				✓

Table B1 shows variance components for the school-by-year random effects model in panel A and the principal and school random effects model in panel B. We show four specifications, beginning with an empty model and successively adding sets of controls. Panel A shows that the magnitude of the school-by-year random effect accounts for roughly 5–10% of the variation in residualized test scores. Panel B shows that even controlling for prior test scores, student characteristics, and school-by-year means of the student characteristics, there remains a substantial unobserved contribution of schools to the total variation in student test scores. Further, the magnitude of the principal and school random effects are roughly equal, underscoring the potential for bias in principal VA estimates that do not include school fixed effects.

Appendix C
Principal Labor Market Simulation

To simulate the process of principal mobility among schools that produces the variation required to estimate principal value-added, we construct a 10-year panel dataset with 1000 schools. These 1000 schools are assigned to 64 districts of varying sizes: 1 district with 300 schools, 1 district with 100 schools, 2 with 50, 5 with 25, 20 with 10, 35 with 5. In year 1 (i.e., the first year of simulation), each school is assigned one principal. According to the processes described below, 20% of principals leave their positions each year. This process occurs iteratively, such that the market is cleared for year 1, which determines year 2 positions, then year 2 separation occurs, etc. The end result is a 10-year panel dataset of principal-to-school assignments. We then merge this to a student-level dataset where each school-by-year has 150 students. Each school has three grade levels (e.g., 3rd, 4th, 5th), where students exit after the last grade. Thus, in each successive year the oldest cohort exits the school and they are replaced by new cohort of 50 students. Across the 10-year simulation, then, each school has 600 students total. For simplicity, we do not model the movement of students across schools.

The key purpose of the labor market simulation is to understand the processes that produce connected networks of principals and schools and their implications for the validity/reliability of principal VA from different estimators. Thus, we produce different simulated datasets by varying the parameters that drive the labor market transitions, keeping fixed the separation rate (20% of principals leave their positions each year). Specifically, we vary four parameters: (1) the proportion of separations that are attrition (i.e., the principal leaves the dataset) versus transfer, (2) the degree of geographic localization among transferring principals, (3) the degree of absolute sorting of principals to schools based on quality, (4) the degree of endogenous separations based on principal quality and the quality of the match between principal and school. We outline each of these parameters below.

Attrition

Attrition is the percentage of principals—among the 20% who leave their positions—that exit the dataset in a given year. This reflects principals who retire, earn promotion to central office, are demoted to a lower school-level position, or who leave the state's public education system for another reason (e.g., to work in a private school). Prior work demonstrates that attrition rates are high—comprising between half and three-quarters of separations (e.g., Grissom & Bartanen, 2019; Grissom et al., 2019). Thus, it is useful to contrast this with a labor market where there is no attrition (all principals who leave their positions move to a different school). We construct datasets where attrition is **high (60%)** and **zero**. Higher attrition rates will tend to limit the size of connected networks and erode their connectivity as fewer principals are observed in multiple schools.

Geographic Localization

Geographic localization captures the extent to which transferring principals are more likely to move among schools that are close to one another. Education labor markets

are highly localized and most principals transfer to schools in the same district. In Tennessee, for instance, only 15% of transfers are across districts. We model geographic localization by specifying the percentage of transferring principals who *must* remain in the same district: **0%** (no enforced localization) and **95%**. Note that in the case of 0%, it is still the case that principals will move within districts due to random chance. To build this into the simulation, we split the market clearing process into two stages. In stage 1, each district fills openings with other transferring principals from the same district (according to the localization parameter). Remaining schools without a new principal are moved into stage 2, where all remaining vacancies are filled using across-district movers and, depending on the attrition parameter, new-to-dataset principals. Higher geographic localization will tend to produce smaller connected networks, all else equal, though these smaller networks may be more strongly connected.

Principal-to-School Sorting

High-quality principals may be more (or less) likely to work in high-quality schools. We simulate this by jointly drawing principal and school effects in year 1 with a correlation of: **-0.4**, **0**, or **0.4**. However, we also need to account for sorting in the subsequent market clearing for each year. To do this, we introduce a parameter whereby the lists of school vacancies and principals are sorted such that the observed correlation of principal and school quality roughly matches the correlations specified for year 1. All else equal, it is unclear how this sorting will affect network sizes or connectivity.

Endogenous Separations

Prior work suggests that higher-quality principals and principals judged more favorably by their teachers are less likely to leave their positions, which suggests that separations are not purely random. To model this, we introduce a weighting parameter that creates a correlation between the likelihood of separation and the simple average of (1) principal quality and (2) match quality of the principal-school pairing. The weighting parameter is set to produce a correlation between turnover (as a binary indicator) and quality (principal + match) of **0** or **-0.3**. All else equal, it is unclear how this nonrandom separation will affect network sizes or connectivity.

Simulation Procedures

For each combination of the four parameters described above (attrition, geographic localization, sorting, endogenous separations), we simulate five datasets. There are 24 combinations, so 120 datasets total. For each dataset, we then run the main simulation 16 times, with eight iterations for a no match effect scenario and eight iterations for a large match scenario. Given computational costs and the fact that the small match case is always bounded between the two extremes, we do not run the small match scenario on these simulated datasets. This yields 1,920 individual simulations, which we analyze in the same manner as the simulations using real datasets.

To summarize the results, we provide a series of tables that results from OLS models regressing the indicated performance measure (correlation with true principal

effect, bias measure, SD ratio) on the various labor market simulation parameters. In additional exploratory analyses, we did not find any meaningful interactions among these parameters. As explained in the main text, we do not present simulations that manipulate the endogenous separations or geographic localization parameters, as the tables below show that they do not greatly impact the performance of the VA estimators.

Table C1

Results for Simulation Mobility Datasets (Correlation Between Estimated and True Principal Effect)

	P+S FE	P+S RE	P-S+ S FE	P+S+ M RE	P FE Only
Panel A: Local Effectiveness					
Base (no match, low att, no loc, no sort, exog sep)	0.934	0.933	0.889	0.935	0.789
Large Match	-0.200	-0.122	0.007	-0.052	-0.031
High Attrition	-0.235	-0.055	-0.001	-0.021	-0.062
Highly Localized Transfers	-0.045	0.000	0.004	0.002	0.005
Negative Sorting	0.007	-0.009	0.005	-0.012	-0.099
Posistive Sorting	-0.009	-0.000	-0.024	-0.010	0.099
Endogenous Separations	0.010	-0.005	-0.010	-0.007	-0.005
Panel B: Global Effectiveness					
Base (no match, low att, no loc, no sort, exog sep)	0.931	0.932	0.721	0.760	0.789
Large Match	-0.192	-0.116	0.026	0.037	-0.031
High Attrition	-0.271	-0.105	-0.006	-0.017	-0.062
Highly Localized Transfers	-0.049	-0.007	0.005	0.005	0.005
Negative Sorting	0.005	-0.011	-0.014	-0.044	-0.099
Posistive Sorting	-0.014	-0.008	-0.063	-0.035	0.099
Endogenous Separations	0.012	-0.001	0.009	0.006	-0.005

Table C2
Results for Simulation Mobility Datasets (Bias Measure)

	P+S FE	P+S RE	P-S+ S FE	P+S+ M RE	P FE Only
Panel A: Local Effectiveness					
Base (no match, low att, no loc, no sort, exog sep)	0.915	0.857	0.948	0.899	0.874
Large Match	-0.045	-0.032	0.001	-0.023	-0.010
High Attrition	0.045	-0.034	-0.003	-0.024	0.047
Highly Localized Transfers	-0.003	-0.000	0.000	0.002	-0.001
Negative Sorting	-0.004	-0.019	-0.004	0.002	-0.236
Posistive Sorting	0.003	-0.006	0.001	-0.015	0.302
Endogenous Separations	0.012	0.004	0.007	0.002	-0.006
Panel B: Global Effectiveness					
Base (no match, low att, no loc, no sort, exog sep)	0.918	0.859	0.624	0.593	0.874
Large Match	-0.039	-0.029	0.036	0.064	-0.010
High Attrition	-0.068	-0.129	-0.011	-0.019	0.047
Highly Localized Transfers	-0.018	-0.014	0.003	0.005	-0.001
Negative Sorting	-0.009	-0.022	-0.035	-0.040	-0.236
Posistive Sorting	-0.014	-0.018	-0.076	-0.067	0.302
Endogenous Separations	0.020	0.011	0.034	0.024	-0.006

Table C3
Results for Simulation Mobility Datasets (SD Ratio)

	P+S FE	P+S RE	P-S+ S FE	P+S+ M RE	P FE Only
Panel A: Local Effectiveness					
Base (no match, low att, no loc, no sort, exog sep)	0.884	0.916	1.066	0.961	1.103
Large Match	0.374	0.105	-0.007	0.032	0.042
High Attrition	0.571	0.024	-0.002	-0.003	0.160
Highly Localized Transfers	0.101	-0.001	-0.004	0.000	-0.007
Negative Sorting	-0.020	-0.011	-0.011	0.016	-0.180
Posistive Sorting	0.020	-0.008	0.031	-0.006	0.217
Endogenous Separations	-0.020	0.011	0.020	0.009	-0.001
Panel B: Global Effectiveness					
Base (no match, low att, no loc, no sort, exog sep)	0.871	0.889	0.864	0.779	1.075
Large Match	0.212	-0.010	0.020	0.049	-0.067
High Attrition	0.538	0.035	-0.009	-0.009	0.220
Highly Localized Transfers	0.080	-0.007	-0.001	0.002	-0.004
Negative Sorting	-0.021	-0.011	-0.031	-0.007	-0.172
Posistive Sorting	0.008	-0.013	-0.030	-0.054	0.210
Endogenous Separations	-0.029	-0.001	0.037	0.026	-0.008

Appendix D
Comparison of Shrinkage Estimators

As described in the "Models For Estimating Principal VA" section, we considered two methods for implementing a shrinkage estimator of principal VA: (1) a post-hoc adjustment to the estimated principal FE from the P+S FE model (shrunken FE) and a random effects approach that yields best linear unbiased predictions (BLUPs) of principals' effects. Our main simulation results include the random effects estimator. We do not include the shrunken FE estimates both because of the extreme computational demands for calculating the standard errors and because the random effects approach has substantially better performance. To demonstrate the latter, we present simulation results from a subset of our data, which allows us to compare shrinkage estimators in a computationally feasible manner. Specifically, we restrict the analysis to the largest connected network in Tennessee, which is a useful test case because it contains the most estimation error and thus should benefit the most from a shrinkage approach.

Appendix Figure D1 shows the simulation results for the FE and two EB approaches. Panel A shows that the P+S RE model substantially improves upon the estimates from the P+S FE model, while the shrunken FE approach does not. For example, the P+S estimates in the no match scenario are correlated with the true principal effects at roughly 0.5, which increases only marginally using shrunken FE but to roughly 0.85 for the BLUPs from the RE model. In the P FE models, there is virtually no difference between the FE estimates and either of the shrunken estimates. As shown in panel C, the RE model also produces a remarkably accurate estimate of the magnitude of principal effects, even in the presence of a large principal–school match component. By contrast, the shrunken FE still overstates the SD of principal effects, particularly when match effects exist.

The differences between the two shrinkage approaches stem from the fact that, in the shrunken FE model, the school effects are included as covariates and are estimated via the FE estimator versus the RE estimator. While FE produces consistent estimates of the school effects, these estimates are plagued by the same estimation error as the principal effects. In essence, the shrunken FE approach does little to address the root cause of the estimation error, which is the collinearity of principal and school assignment. The RE model, which assumes (incorrectly) that principal and school assignments are uncorrelated, produces biased, yet substantially more precise, estimates of the principal and school effects. The net result is that the bias/variance tradeoff is squarely in favor of the RE approach when the source of estimation error is the collinearity of the principal and school effects (i.e., the P+S model). In the P FE case, however, there is little estimation error because there are not school FE. Here, the two shrinkage approaches are more or less identical. With the large number of students contributing to each principal's estimated effect, the FE and shrinkage estimates are very similar.

Comparing Panels B and C demonstrates the bias/variance tradeoff more clearly. In the P+S model, both shrinkage procedures reduce estimation error, which yields a SD of principal VA close to the true SD of principal effects. This shrinkage, however, results in biased estimates, whereby differences in shrunken VA understate the true differences in principal effectiveness. In the case of shrunken FE, the gain in precision is almost perfectly offset by the bias, while the bias in the mixed model is much smaller. In the P FE model,

the FE estimates are already very precise, such that there is little to be gained through shrinkage.

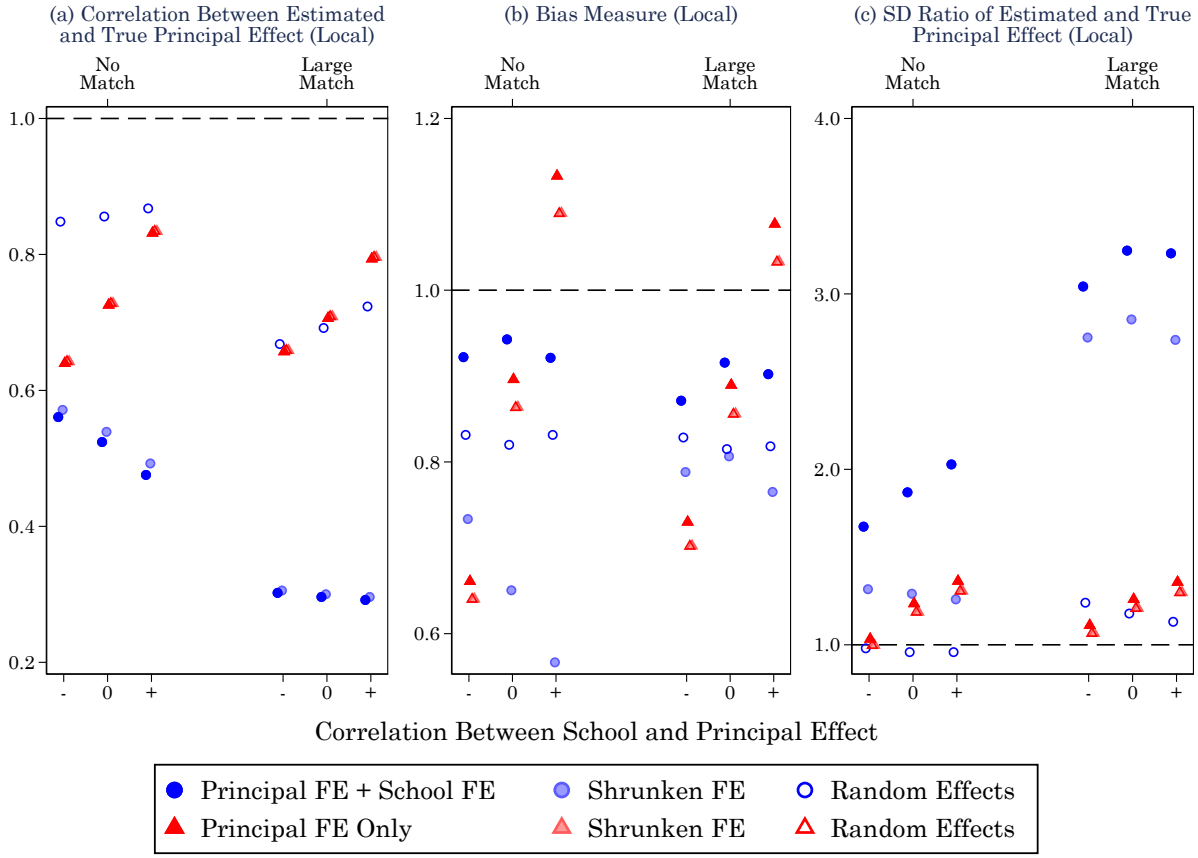


Figure D1

Comparison of Shrinkage Estimators

Notes: Shrinkage analysis only performed for principals in the largest connected network in Tennessee. In each plot, the y-axis is defined by the header. The top y-axis corresponds to the magnitude of the principal–school match effect (0%, 22%, 56% of the total principal effect). Additional horizontal spacing is to facilitate visual comparisons between models. In the results shown here, the school performance effect constitutes 5% of the total variance in student test score growth. In Panels A, B, and C, the true principal effects are residualized on network fixed effects corresponding to the connected networks formed by principals and schools for the given VA measure. For ease of interpretation, a small amount of horizontal spacing is added between models.