Are Schools More than a Collection of Teachers? Joint Estimation of Teacher and School Value-Added*

Jesse Bruhn Isaac M. Opper Sai Pitre

March 26, 2026

PRELIMINARY AND INCOMPLETE. DO NOT CITE OR DISTRIBUTE.

Abstract

Value-added models typically attribute student test score gains to a student's teachers or their schools, but teachers and schools may both play a role in determining student outcomes. In this paper, we first provide reduced-form empirical evidence that accounting for school quality is important for correctly estimating teacher quality (and vice versa). We then develop a new statistical approach to jointly estimate teacher and school value-added. In doing so, we find that schools differ in how well they educate students, even after accounting for varying levels of teacher quality at the schools. This implies that ignoring school effects biases conventional measures of teacher value-added, which has important implications for both the theoretical understanding of educational production and practical decisions related to the day-to-day administration of schools and districts. For example, we find (1) that the tendency for teachers to move to better schools early in their career implies that standard value-added methods overstate the returns to experience by 33%; and (2) that even with infinitely many years of data, status quo-methods will incorrectly rank teachers in at least 20% of cases and, on average, yield gains that are 50% smaller than would otherwise be expected.

*We thank Andrew McEachin, Christopher Candelaria, Christopher Doss, Eric Hedberg and Eric Taylor for many helpful discussions on how jointly estimate teacher and school effectiveness. Susha Roy and Ini Umosen both made valuable contributions to the initial analysis that led to this paper.

I Introduction

Test-score based accountability is a high-stakes statistical exercise. As of 2024, at least 11,500 school districts, employing over 2.3 million educators and responsible for the well-being and academic development of over 34 million children, used some form of growth-based measure (like value-add) in their formal school and/or teacher evaluation systems.[1] Thus it is critical to have reliable measures of value-added: policymakers use these methods to help determine which teachers are hired, fired, and promoted, and which schools should enter receivership in order to be reformed or even permanently closed (e.g. Bates et al., 2025; Abdulkadiroglu et al., 2026).

In practice, estimating value-add requires thorny choices about who should obtain credit, in a statistical sense, for student test score growth. Existing approaches typically treat schools as no more than collections of teachers–at best, treating school effects as nuisance parameters that must be accounted for when estimating teacher value-added (e.g. Bruhn et al., 2022; Eastmond et al., 2026; Bates et al., 2025)–while many others ignore the issue altogether, making no attempt to distinguish school effects from the underlying teacher workforce. Yet, this implicit assumption contradicts decades of high-quality empirical research on the importance of independent school level inputs, like tutors and guidance counselors. Thus, existing methods may be biased in ways that shape our understanding of the education system and lead to sub-optimal policy decisions.[2]

In this paper, we study the joint distribution of teacher and school value-added. We make three contributions. Our first contribution is to provide reduced form evidence that suggests schools are more than just a collection of teachers. This motivates the need for better statistical methods of apportioning test score growth between schools and their teacher

[1] Calculations are the authors' own based on data from Swisher and Saenz-Armstrong (2022) and U.S. Department of Education, National Center for Education Statistics (2023).

[2] This tension is more than just a theoretical concern. Many states and districts have taken a decidedly "ad-hoc" approach to the way they handle the between-school component of test scores growth. In Florida, for example, the official statistical model used for high stakes evaluation "compromises" on this question by arbitrarily giving 50% of school value-added to each of the individual teachers that work there, a tacit acknowledgment by policy makers that our current understanding of the separate, statistical roles of teachers and schools is about as certain as the flip of a coin (Florida Department of Education, 2015).

labor force. Our second, and primary, contribution is to develop new tools for identifying the joint distribution of teacher and school value-add using the network of schools generated by teacher switching. Our method is similar to those developed in the literature on worker-firm fixed effects (Abowd et al., 1999; Card et al., 2013; Kline et al., 2020), but allows for less stringent identifying assumptions and weak network ties. Our third and final contribution is to demonstrate that separate identification of the teacher and school components of value-add has important implications both for the real world educational policies they inform, and for our theoretical understanding of the nature of educational production.

The setting is North Carolina. Our sample consists of the universe of public school students, and their teachers, in grades 3-8 as observed over the years 2015-2019. The final sample we use for estimation consists of 1.2 million students and 42,000 teachers. Importantly, many of the teachers in our sample ($\approx 20\%$) switch schools at least once over our sample frame. This variation is key, since it will allow us to explore, empirically, the existence of independent school effects. However, exploiting this variation is complicated by the fact that teacher moves are non-random. While teacher level differences between movers and stayers on observables like race, ethnicity, age, and years of experience are modest in size, we *do* find large differences in the types of schools teachers tend to switch between. This pattern is well summarized by test scores: on average, students in the receiving school have 10-15% of a standard deviation larger test scores than those in the school the teacher leaves behind, a fact that will be important for our identification strategy.

To begin, we offer a novel definition of independent school and teacher value-add that is anchored in potential outcomes. Defining independent school and teacher effects is less straightforward than it might seem, since it requires us to consider complicated edge cases. For example, if a given school is systematically better than others at finding, hiring, and retaining high value-added teachers, then should the value-add of the school reflect the improvement in teacher value-add induced by better personnel management? Similarly, what if some schools are better than others at "growing" highly effective teachers via professional development? Intuitively, our definition addresses these edge cases by partitioning test score

growth according to its "persistence" across hypothetical teacher-school assignment configurations. Under this definition, schools will not *directly* receive credit for better personnel management. However, since our approach characterizes a *joint* distribution of teacher and school value-added, it is trivial to examine directly which schools are able to systematically hire (or develop) high value-added teachers and reward them appropriately, something that is impossible to do quantitatively using current methods that examine teachers and schools in isolation.

Our first finding is that schools are more than a collection of teachers. To support this claim, we offer two pieces of evidence. First, we show that existing tests of forecast bias that use teacher transitions to assess the validity of value-added models primarily use *within* school variation and hence are unable to detect biasses, like those caused by an independent school effect, that only emerge across schools. Versions of this test that we explicitly modify to detect cross-school sources of bias definitively reject that existing value-added procedures are forecast unbiased. Second, we show that schools systematically predict changes in value-add among switchers in a manner that is consistent with independent school effects. Specifically, we find that teachers who make opposite switches, such that the origin-destination pair of teacher A become the destination-origin pair of teacher B, experience changes in teacher value-added that are equal and opposite in sign around the timing of the move. This suggests that conventional value-added measures are being pushed up (or pulled down) by the independent value-add of the school where they choose to work.

Motivated by this evidence, we develop a novel statistical method in the spirit of Abowd et al. (1999) and Card et al. (2013) that leverages teacher switching to separately identify teacher and school components of student test score growth. Motivated by our evidence that teachers who switch schools ("movers") are observationally different than teachers who do not ("stayers") and that both the timing and direction of teacher transitions are non-random, our method relies only on comparing movers to other moves and allows for the timing of the move to be endogenous. We use this method to show that the is indeed variation across schools – conditional on teacher quality at the school – in test score value-

added. This variation, however, is less than the variation across teachers in test score value-added and there is very limited sorting of teachers across schools based on their value-added. We also highlight how we can use these estimates as hyperparameters in an empirical Bayes model that outputs estimates of both individual teacher and school quality.

Our third, and final, finding is that separately identifying school and teacher value-add has important implications both for educational evaluation systems and our theoretical understanding of educational production. On the theoretical side, for example, we find that status-quo methods of estimating teacher value-added over-state the returns to experience by ≈33%. This results directly from the fact that teachers tend to move to better schools early in their career—an empirical pattern that is impossible to detect without estimates of the joint distribution of teacher and school value-add. On the policy side, we show that even with infinite data, a district administrator using value-added estimates to compare teachers across schools—as is done in systems that rank teachers on a common district-wide scale (e.g., Washington, D.C.'s IMPACT system; District of Columbia Public Schools, 2010; Dee and Wyckoff, 2015)—would draw incorrect inferences in at least 20% of cases and, on average, yield gains that are 50% smaller than would otherwise be expected. By way of contrast, our method yields correct inference in 90% of cases with only two years of data, and nearly 100% of cases with 7 years of data, which demonstrates its practical utility for real world policy decisions.

II Related Literature

We contribute to the literature on teacher value-add by developing a novel method for jointly identifying school and teacher effects. A large literature studies whether conventional teacher value-added measures recover teachers' causal contributions to student achievement and how best to estimate those contributions in practice, with particular emphasis on student sorting, forecast bias, validation under random assignment, and finite-sample shrinkage (e.g., Rockoff, 2004; Kane and Staiger, 2008; Rothstein, 2010; Koedel and Betts,

2011; Chetty et al., 2014; Kane et al., 2013).[3] Of particular note are Mansfield (2015) and Jackson (2013). Mansfield (2015) is the first paper that recognizes the potential for school effects to bias estimates of teacher value-add. He proposes to address this issue by estimating models with school fixed effects, effectively treating the schools themselves as a nuisance parameter rather than an object of independent interest.[4] Jackson (2013) studies match effects between teachers and schools, leveraging the network of switchers for identification. We build on these papers by (1) recognizing the potential for weak network ties to create important forms of bias, (2) treating the school effects themselves as a substantive object of interest; and (3) developing novel methods to identify both a joint distribution of teachers and schools quality and to use this to estimate individual teacher and school quality for use in practice. Thus, relative to the existing body of work, our contribution is not simply to further refine estimation of a single teacher parameter. Instead, we complement the existing teacher value-added literature by recovering the teacher effect relevant for across-school comparisons while simultaneously estimating school value-added as an object of independent economic interest.

We also contribute to the literature on the identification of worker and firm fixed effects by developing a method that both imposes less stringent identifying assumptions and that addresses weakly tied firm networks using the techniques of empirical-Bayes. Beginning with Abowd et al. (1999), a large literature has used linked worker-firm data and additive two-way fixed-effects models to decompose outcomes into worker and firm components and to study sorting and firm premia (e.g., Card et al., 2013). A parallel methodological literature emphasizes that identification in these models is fundamentally a network problem: limited mobility can distort estimated sorting patterns, weak connectivity can make firm effects imprecisely estimated, and variance decompositions require careful bias correction (Andrews et al., 2008; Jochmans and Weidner, 2019; Kline et al., 2020). Recent work has responded

[3]For an overview, see Koedel and Rockoff (2015a). For more recent methodological work, see Gilraine et al. (2020); Gilraine and Pope (2021); Rose et al. (2022); Mulhern and Opper (2021); Eastmond et al. (2026), and Tartova (2023).

[4]Several recent papers also build on this general approach by giving more attention to the structure of the underlying network (e.g. Eastmond et al., 2026; Bates et al., 2025).

by using empirical-Bayes shrinkage to stabilize two-way estimates under limited mobility (Cheng et al., 2025), while other papers relax strict additivity by allowing worker-firm interactions or match effects (Woodcock, 2015; Bonhomme et al., 2019). In education, these concerns are especially salient: networks tend to be sparse and highly leveraged (Bartanen and Husain, 2022); moreover, recent experimental evidence shows that teacher effectiveness is not fully portable across school contexts (Kraft et al., 2026). Relative to this literature, our contribution is to combine these insights in a unified framework: we relax the assumptions of full portability and exogeneity of transitions that are embedded in canonical worker-firm decompositions and then use empirical-Bayes methods to recover informative teacher and school effects even in weakly connected teacher-school networks.

III Conceptual Framework

III.A A potential outcomes definition of teacher and school value-add

An important first step is to precisely define teacher and school value-add such that they can be considered collectively. Doing so is challenging, however, because it requires researchers to take a stand on difficult conceptual questions about whether teachers or schools should "obtain credit," in a statistical sense, for different types of educational inputs. Some educational inputs are easy to classify. Clearly, extensive tutoring support that is offered by the school outside of the classroom environment is a feature of the school itself. Similarly, the teacher should receive credit for the innate ability or drive she brings to the job. However, many other educational inputs are less cut and dry. For example, suppose that some schools are systematically better than others at recruiting and retaining teachers with high-value add. Should the value-add of this school include the improvement in teacher value-add induced by this better personnel management? Should the teachers get credit? Or both? Similarly, suppose that some schools offer excellent professional development, which in turn improves the value-added of teachers that work there. Should the test score gains from this intervention accrue to the teacher or the school?

We address this challenge by offering a simple, statistical definition of value-add anchored in potential outcomes. Intuitively, our definition will partition growth according to its "persistence". For a school, persistence is the part of growth robust to replacing its underlying labor force with an average one . For a teacher, persistence refers to the component of observed growth that would "disappear" if she were replaced by a random teacher.

The building block of the model will be "teacher-by-school" effects, i.e., the combined value-add of the teacher and school in a specific year. We will denote this as r^*_{jst} for teacher j in school s in year t. Note that, as is generally the case when using a potential outcome framework, r^*_{tjs} is defined for every potential school and teacher combination (in year t), even though the teacher is only observed in a single school in year t. Note that this structure allows teacher j to be particularly effective in school s, for example, either because of the specific mix of students at this school or because the additional supports that the school offers are particularly complementary to her teaching style.[5]

We can then define teacher value-added (TVA) as what would happen (in expectation) if teacher j got replaced at school s in year t by a random teacher. And we can define school value-added (SVA) as the effect of school s in year t on their students if a random set of teachers were assigned to that school. Formally, that

$$\theta_{jst} = r^*_{jst} - |J|^{-1} \sum_{\forall j} r^*_{jst} \tag{1}$$

$$\mu_{st} = |J|^{-1} \sum_{\forall j} r^*_{jst} \tag{2}$$

where θ_{jst} is teacher j's TVA in year t (if she taught in school s) and μ_{st} is school s's SVA in year t. Note that $r^*_{jst} = \theta_{jst} + \mu_{st}$, so the definitions of TVA and SVA cleanly partition the potential test score gains regardless of how teachers get assigned to schools.

A key advantage of this definition is that it allows transparent, richer, and more nuanced approaches to the question of who should get credit for which kinds of educational inputs.

[5]This formulation does implicitly assume away the influence of specific student-to-teacher matches *within schools*. This is consistent with the focus of this paper, which is on across-school sorting.

For example, according to this definition, schools will not directly receive credit for their ability to attract and retain a persistently high value-add labor force. However, because this definition characterizes a *joint* distribution of teacher and school quality, it would be simple to use these parameters to ask whether school A has higher average teacher value-add than school B, even if this does not appear directly in the school value-add measure.

We acknowledge, however, that these definitions of TVA and SVA are asymmetric, in that they "assign" any potential match effects to the teacher rather than the school. Conceptually, this is the natural way to partition the match component for a decision maker like a school principle who, for personnel policy like firing, is likely to care most about the realized influence of a teacher at the school where they currently teach, and should care less about the "hypothetical" impact one of their employees could have if they left for a different school. Furthermore, even in the presence of sizable match effects, we would argue that value-added methods aligned with this definition hold the potential to represent a marked improvement relative to existing models, a point we make clear in the next section.

Throughout, we will allow TVA to drift from year-to-year, but assume that SVA is static and so generally denote it as μ_s. We will also use θ_{jt} to denote the TVA of teacher j in year t in the school they taught that year, which is usually the object of interest. In principle, we could define school-value add at the school-by-year level similar to the teacher definition; in practice, schools are large organizations that are slow to change and so we expect this approximation, which will improve precision considerably in our empirical application, to be reasonable.

III.B What does the existing approach to value-add identify?

Existing approaches to value-added estimation consider teachers and schools in isolation. This implicitly gives teachers credit for the entirety of school-level inputs, and it gives schools credit for the average value-add of their teacher labor force. To see why, observe that existing approaches to identification of teacher and school value-add are, up to empirical-bayes

adjustments meant to reduce the influence of sampling variation, equivalent to separately averaging residual test scores over observed student-teacher, or student-school, assignments.

For simplicity, assume that there is no sampling variation and that we continue to focus on TVA and SVA estimates for a single year t. In that case, standard TVA models identify:

$$\begin{aligned} \hat{\theta}_{jt} &= r^*_{jst} \\ &= \theta_{jt} + \mu_{st} \end{aligned} \tag{3}$$

and standard SVA models identify:

$$\begin{aligned} \hat{\mu}_{st} &= |J_{st}|^{-1} \sum_{j \in J_{st}} r^*_{jst} \\ &= \mu_s + |J_{st}|^{-1} \sum_{j \in J_{st}} \theta_{jt} \end{aligned} \tag{4}$$

where J_{st} denotes the set of teachers that actually work in school s in year t. Thus, the current approach could lead to consequential errors as a result of confounding. In particular, while it is unclear how one should handle the match effects, few would argue that one should include SVA (i.e., μ_s) in the measure of TVA. If schools vary widely in their effectiveness, for example, conventional measures would likely lead policy makers to mistakenly label many high value-add teachers as low performing, and many low value-add teachers as high-performing, simply because of the out-of-class resources available in the school where they have chosen to work.[6]

In recognition of this potential for confounding, it is increasingly common in applied work for researchers to avoid this problem by estimating teacher value-added entirely within-school (e.g. Bruhn et al., 2022). However, this is limiting because it only identifies the *relative* value-added of the teachers assigned to work in that school, without the ability to

[6]Note that how misleading the TVA measures are about a teachers' rank depends on both the amount of dispersion in the SVA measures and their correlation with TVA.

compare them to other peers who do not work at the same school. While this may be a useful metric for some policy decisions, it is not useful for many others, like district-wide personnel management.[7] Ultimately, whether it makes sense to fire a teacher or deny them professional status depends critically on the expected value-add of the replacement teacher, not just the current teacher's relative position in the within-school distribution.

If school effects are small, it is theoretically possible that existing approaches could be "good enough," for most practical decisions. For example, if schools have no effect, as is implicitly assumed by many current TVA models, then there is truly no meaningful distinction between schools and and their teacher labor force. In that case, schools really are nothing more than a collection of teachers, and the average quality of the teacher labor force at a given school is a sufficient statistic for its quality. Mathematically, we see that if $\mu_s \approx 0$, then standard teacher value add identifies true TVA and standard school value add identifies the average teacher quality at that school. However, there is now a large body of work documenting the importance of school-level inputs such as guidance counselors, tutors, and community school initiatives (e.g., Mulhern (2023), Cortes et al. (2024), and Covelli et al. (2024), respectively). Therefore, it cannot be the case that schools are just a collection of teachers, as this would contradict decades of high-quality, empirical work. In our view, whether school effects are "small enough" for existing models to work well is therefore, ultimately, an empirical question. We next describe our data and then provide some empirical evidence that suggests they are not.

IV Data

We use anonymized administrative data from the North Carolina Department of Public Instruction (NC DPI). The data contain information on every student in grades 3-8 at

[7]This is not a theoretical concern. Consider the controversial firing of Sarah Wysocki, a fifth grade elementary school teacher in Washington, D.C., who was fired *against* the wishes of her principal as a result of a requirement in the district's IMPACT evaluation system that mandated teachers be let go after two years of low-growth, where low-growth was defined relative to the district as a whole (Turque, 2012).

a public school in North Carolina from the 2015 school year until the 2019 school year.[8] The data contain yearly information about each student's grade-level, school attended, and assigned math and English teachers. They also contain student demographic information, including gender, race and ethnicity, whether the student has Limited English Proficiency (LEP), whether the student is economically disadvantaged, and whether the student has been identified as being academically gifted. We also observe students' end-of-grade math and English (ELA) test scores. Because tests change each year, we follow convention by normalizing these scores by subject, grade, and year to have a mean of zero and a standard deviation of one. The analysis uses data from 2015-2019 to estimate teacher and school effects in 2016-2019 subject to a number of sample restrictions which are standard in the teacher value added literature.[9]

The analytic sample consists of 1,200,000 students and 42,000 teachers in public elementary and middle schools in North Carolina from 2015 to 2019. The student population is 48% white, 25% Black, and 18% Hispanic. Teachers are in the data for 2.75 years, on average, and almost a third of teachers appear only once during the 5 year panel. In the average year, 5 percent of teachers change schools, 71 percent stay at their school, and 23 percent enter the data for the first time. There is very little variation in these statistics from year to year.

We use the analytic sample to calculate the average teacher-level residuals, which we denote as r_{jt}, for teacher j in year t. To do so, we first estimate the following regression model:

$$y_{it} = \beta X_{it} + \nu_{j(i,t)} + u_{it} \tag{5}$$

where y_{it} are test scores for student i in year t; X_{it} are a collection of predetermined (as of time t) control variables; $j = j(i,t)$ is an index that describes the teacher j assigned to student i in year t; and ν_j is a teacher fixed effect. Control variables include student lagged

[8] We follow the convention of referring to academic years using the spring semester, e.g., the 2015 school year refers to the 2014-2015 academic year.

[9] The restrictions remove about 14% of total observations. See Appendix B for more detail.

test scores, gender, race/ethnicity, and indicators for their LEP status, disability status, whether they are economically disadvantaged or academically gifted. We then calculate student-level residuals as $y_{it} - \hat{\beta}X_{it}$ and subsequently aggregate them to the teacher-year level to get:[10]

$$r_{jt} = |I_{jt}|^{-1} \sum_{i \in I_{jt}} y_{it} - \hat{\beta}X_{it} \tag{6}$$

where I_{jt} is the set of students taught by teacher j in year t. The residuals measure how much better (or worse) a teacher's students perform in math or ELA tests than would be expected based on lagged test scores and demographic characteristics. These average residuals are the observed empirical analogue to the r^*_{jst} from the potential outcomes model in the preceding section and hence form the main unit of analysis for our empirical work. More precisely, $r_{jt} = \sum_{s' \in S} 1(s(j,t) = s') r^*_{js't} + \epsilon_{jt} = r^*_{jst} + \epsilon_{jt}$, where $s = s(j,t)$ is an index that denotes the school s teacher j worked for during year t; and ϵ_{jt} is a mean zero, normally distributed error (e.g. from the "pre" steps necessary to estimate r_{jt} from the average residuals). Note that, this formulation implicitly assumes that the value-add model chosen by the policy maker or researcher is "correct" in the sense that the r^*_{jst} implied by the choice of covariates, etc. to include in model 5 are, indeed, the objects of interest relevant for whatever decision or analysis is at hand.[11]

Combining this with Equation (3), we get that:

$$r_{jt} = \theta_{jt} + \mu_{s(j,t)} + \epsilon_{jt} \tag{7}$$

i.e., that the measured average residual can deviate from the teachers' true TVA (θ_{jt})

[10] In practice, we only include such residuals for teachers who taught more than 10 students in a given year.

[11] The proper choice of covariates to include in a value-added model has a long history of debate in the economics of education (e.g. Rothstein, 2010; Kane and Staiger, 2008; Chetty et al., 2014; Koedel and Betts, 2011; Kinsler, 2012). The best empirical evidence suggest that the inclusion of lagged test scores is sufficient in order for r^*_{jst} to represent the causal effect that teacher j had on their students in year t, relative to an average teacher who had zero impact (Koedel and Rockoff, 2015b). Consistent with this, we will interpret our estimates as causal throughout. We pause here only to note that our formulation is more general, in the sense that the framework we develop can be applied to *any* evaluation metric (e.g. causal, predictive, normative, etc.) where independent school effects are important to measure. In that sense, we take the choice of controls as given or otherwise determined by the researcher.

because of school effects (μ_s) and mean zero error (ϵ_{jt}).

While we refer to j as a "teacher," we separately estimate these for each subject and so generate subject-specific teacher residuals. This is because many teachers (namely middle school teachers) do not teach both Math and English and the ones that do may perform differently across subjects. Finally, we estimate everything separately for middle schools and elementary schools to allow for potentially changing relationships between the covariates and the outcomes.[12]

V Motivating Evidence

V.A An across-school test of TVA bias

Existing work on bias in teacher value-add uses one of two approaches. The first approach leverages the explicit random assignment of teachers to classrooms within a school in a randomized experiment, the largest of example of which is the Measuring Effective Teaching project (e.g., Kane et al. (2013)). In effect, these tests ask, "Does the difference in value-add predict the causal effect of the random assignment?" However, because these tests rely exclusively on within school variation, any school level effects will necessarily be differenced out. For that reason, these tests only measure whether teacher value-add is unbiased *within* school; not whether important sources of bias exist across schools.

The second test deployed in prior literature is based on teacher transitions (e.g. Chetty et al., 2014). These tests ask whether a 1 standard deviation change in average teacher value-added that emerges from a teacher transition at the school-by-year-by-grade level results in a 1 standard deviation change in student test score outcomes. The intuition is that if teacher value-added is an unbiased measure of teacher effectiveness, then an exogenous change in school-by-grade-level teacher value added emerging from a transition should accurately predict actual, out-of-sample changes in test score residuals.

[12]Note that this means if a teacher switches from teaching at an elementary school to a middle school, we consider them to be different teachers. In contrast, if they switch between two elementary schools (even if they switch grades), we consider them to be the same teacher.

However, in practice, this second test has a similar limitation as those based on random assignment. This is because the definition of a teacher transition typically used for this test includes not only those teachers who switch schools, but also those teachers who switch grades within school or enter the data for the first time. Consider, for example, the case where a teacher who taught 3rd grade in one year who then leaves the district or state and is replaced with a a new teacher (with no history of teaching in the district/state). This is considered a transition for the purposes of the test, since average teacher value-added does change in that school-grade-year; however, the expected school effect remains constant for this transition, since it only leverages within school variation, similar to the randomized experiments described earlier. While these are not the only type of transitions, they constitute a substantial share. For example, Chetty et al. (2014) show that, in an average year in their data, 30.1% of teachers move across grades within school, 5.8% of teachers leave the district, with only 6.1% teacher changing schools. We find similar patterns in our data. Thus, the vast majority of transitions that would be used in this test of forecast bias only leverage within-school variation. For that reason, this test is also more accurately interpreted as showing that the *within-school* sources of bias are minimal.

To be clear, these tests have been incredibly important in our understanding of teacher value-added estimation. For many important research and policy questions, specifically those that only require a within-school distribution of teacher effectiveness, these tests of bias are the appropriate ones. That said, they do not speak directly to whether across-school sources of bias–like those potentially caused by school effects– exist and are of consequential magnitude. To isolate the question of across-school bias, we therefore modify the classic test of forecast bias based on teacher transitions by specifically focusing on teachers who switch schools. This modified test will directly examine biases that arise due to school level differences. The logic is as follows: if school effects *don't* exist, then unbiased estimates of teacher value-added should be equally predictive for teachers who switch within the same school as for those who switch between schools.

To operationalize this, we define two sets of teachers: those who remained within the

same school during the entire sample (the "stayers") and those who switched schools during the same time period (the "movers"). We then generate traditional value-added measures using the same general approach as in Chetty et al. (2014) (henceforth, CFR). We call our measure school-specific teacher value-added (SSTVA), denoted $\hat{\theta}_{j,t}^{SSTVA}$, for reasons that will be clear below. Specifically, we follow CFR in using a jackknife procedure by using residuals $r_{j,t'}$ for $t' \neq t$ to construct the teacher value-added estimates for year t. This ensures there is no mechanical correlation between $\hat{\theta}_{j,t}^{SSTVA}$ and $r_{j,t}$. The resulting SSTVA measures for stayers are identical to the CFR measures, which we calculate as follows:

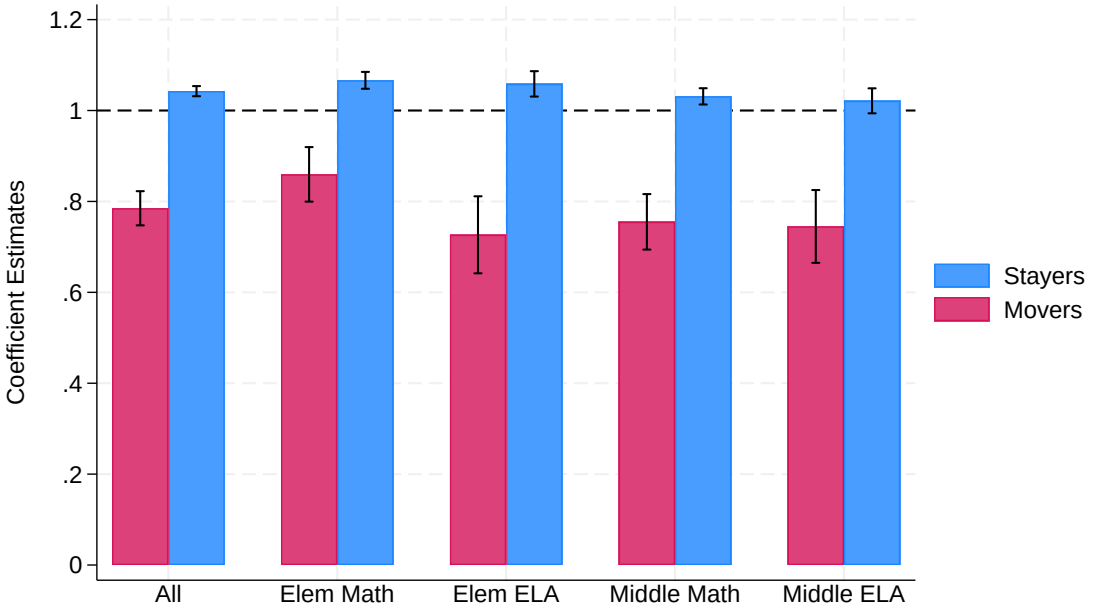
$$\hat{\theta}_{j,t}^{SSTVA} = \sum_{\forall t' \neq t} \psi_{|t'-t|} \cdot r_{j,t'} \tag{8}$$

where the $\psi_{|t'-t|}$ parameters are estimated using the empirical relationship between $r_{j,t'}$ and $r_{j,t}$ in the population. For the movers, the definition of SSTVA is nearly identical; however, we use student residuals from years when a teacher isn't teaching in school s' to estimate $\hat{\theta}_{j,t}^{SSTVA}$. That is, we exclude $r_{j,t}$ and any other year where $s(j,t) = s(j,t')$ when estimating $\hat{\theta}_{j,t}^{SSTVA}$.[13]

We then test whether the resulting SSTVA measures are biased by estimating the relationship between SSTVA and students' test scores separately for movers and stayers. As shown in Figure 1, SSTVA are *less* predictive for teachers who switch schools than for those who stay at the same school. While the coefficient on $\hat{\theta}_{j,t}^{SSTVA}$ is indeed (roughly) 1 for the stayers, the coefficient is significantly less than one for the movers (around $0.7 - 0.8$ for all the samples). This is true in both the pooled sample and the subjects/grade-level combinations. Note that this does not invalidate CFR's mover design and, as we show in Appendix Table 6, we can replicate CFR's mover design using our data. Instead, as discussed above, the differing results stem from the fact that the two tests focus on different sources of bias in the TVA measures.

[13]The $\psi_{|t'-t|}$ weights are the same when estimating both the movers and stayers, and are estimated using a pooled sample.

Figure 1: Forecast Bias for Movers vs Stayers



Notes: This figure implements the test of forecast bias developed in Chetty et al. (2014), except that we split teacher transitions into those that occur across schools (Movers) and those that occur within schools (Stayers). A switcher is defined as in the original paper. Each bar reports the coefficient β from regressions of the form $Y_{ist} = \alpha + \beta\hat{\mu}_{j(s,t)} + \gamma X_{it} + \delta_s + \lambda_t + \varepsilon_{ist}$, where Y_{ist} is student achievement, $\hat{\mu}_{j(s,t)}$ is a leave-year-out jackknife estimate of teacher value-added, X_{it} are student controls, and δ_s and λ_t are school and year fixed effects. Whiskers denote 95% confidence intervals based on standard errors clustered at the teacher level. Samples, denoted on the x-axis, correspond to different subject-by-grade combinations.

V.B A novel test for independent school effects

Motivated by the finding of bias in the preceding section, we develop a new test that allows us to provide evidence regarding the potential magnitude of variation in school effects. While the modified test in the preceding section provides clear evidence of bias in school value-added, it is also not constructive. In other words, it does not provide a clear path for disentangling school level variation from teacher level variation.

To address this, our new test will explore consistency in changes to average student residuals when teachers make the same or opposite moves between schools. Intuitively, teachers who make the same moves, for instance from school s to school s', experience the same change in school effects. As such, if school effects exist, we would expect teachers making the same moves to experience a similar change in their average student residuals when they move. With the same logic, teachers who make the *opposite* moves–one from school s to school s' and the other from school s' to school s–should experience the *opposite* change in school effect.

Formally, if at least two teachers switch from school s to school s', then we have multiple noisy measures of the *differences* in school effects between these two schools, i.e. of $\Delta s = \mu_s - \mu_{s'}$. In particular, we have $\Delta r_{jt} = \Delta\mu + e_{jt}$ and $\Delta r_{j't'} = \Delta\mu + e_{j't'}$ for some error terms e_{jt} and $e_{j't'}$. If we are willing to assume that the error terms are uncorrelated, i.e., that $Cov(e_{jt}, e_{j't'}) = 0$, we can separate out the signal ($\Delta\mu$)–the true difference in school effect–from the noise ($e_{j,t}$ and $e_{j',t'}$), by looking at the relationship between the measures generated by multiple teachers making the same moves.

We first consider teacher j and teacher j' who both move from school s to school s'. We calculate the change in their average student residuals between the years that they move, i.e. $\Delta r_{j,t}$ and $\Delta r_{j',t'}$ and then estimate the following regression:

$$\Delta r_{j,t} = \beta_1 \Delta r_{j't'} + e_{j',t'} \tag{9}$$

The coefficient of interest is β_1. We expect this covariance to be positive, given that

these teachers experience the *same* change in hypothesized school effect. We repeat this exercise for pairs of teachers who make the opposite moves–one from school s to school s' and the other from school s' to school s. In this case, we expect the covariance to be negative, given that the teachers experience the *opposite* change in hypothesized school effect.[14]

Figure 2 depicts this idea visually. We plot in blue the correlation between the change in average student residuals for teacher j and teacher j' who made the same move. The correlation between the change in average student residuals for teacher j and teacher j' who made the *opposite* move is plotted in red. Table 1 provides the regression analogue to Figure 2.

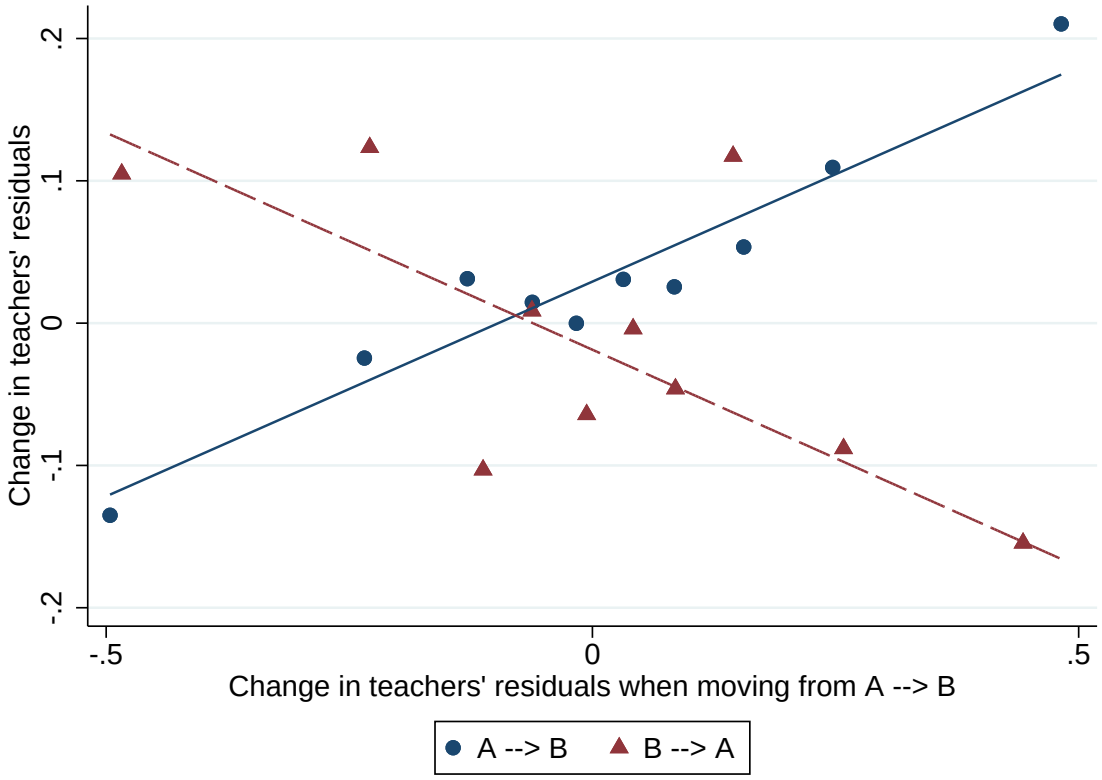
As expected, we find that teachers who make the same move experience positively correlated changes in their average student residuals. Teachers who make the opposite moves experience negatively correlated changes in their average student residuals. Moreover, these relationships (both same and opposite moves) are of a similar magnitude. This pattern is consistent with the existence of an unobserved school effect: average student residuals change in a predictable manner when teachers move between the same schools.

This exercise has an additional advantage: By using all pairs of teachers who switch between the same two schools, we can estimate the variance in the observed difference in school quality. In particular, under the assumption that the error terms are uncorrelated, we have that $Cov\big(\Delta r_{j,t}, \Delta r_{j',t'}\big) = Var(\Delta\mu_s)$; and hence we also use these moves to estimate the standard deviation of the difference in school effects. We find that the implied standard deviations range from 0.10 to 0.20 (See table 1) depending on the subject and grade-level.

This estimate of $Var(\Delta\mu_s)$ identifies the variance of the true difference of school effects under a relatively weak assumption: that the error terms of the two moving teachers are independent. Note, however, that although the estimate of the change in school effects, $Var(\Delta\mu_s)$, may offer insight into the the variance of school effects, $Var(\mu_s)$, the

[14] In practice, we form the teacher pairs $(\Delta r_{jt}, \Delta r_{j't'})$ that feed into these regressions by randomly choosing within origin-destination schools.

Figure 2: Consistency of residual changes when switching schools



Notes: This figure visualizes a test of consistency in changes to average student residuals when teachers move between the same pair of schools. We form pairs of teachers who either make the *same* move (e.g., from school s to s') or *opposite* moves (one from s to s' and the other from s' to s). For each teacher j, we compute the change in average student residuals at the time of the move, Δr_{jt}. We then estimate regressions of the form $\Delta r_{jt} = \beta_1 \Delta r_{j't'} + \varepsilon_{j,t}$, where (j, j') index pairs of teachers making either the same or opposite moves. The figure reports binned scatter plots of Δr_{jt} against $\Delta r_{j't'}$, along with fitted regression lines. Blue circles correspond to teacher pairs making the same move, while red triangles correspond to teacher pairs making opposite moves. Consistent with the presence of school effects, changes in residuals are positively correlated for teachers making the same move and negatively correlated for those making opposite moves. We acknowledge that the estimates for opposite moves (red triangles) visually appear to exhibit more noise, reflecting the smaller sample of teachers who move across opposite origin destination pairs; however, the relationship is precisely estimated. Corresponding regression results, including formal tests of statistical significance, are reported in Table 1.

Table 1: Correlation of Teacher Residual Change after Multiple Transitions

(a) Multiple transitions from School A to School B

	(1) $\Delta r_j, t$	(2) $\Delta r_j, t$	(3) $\Delta r_j, t$	(4) $\Delta r_j, t$
$\Delta r_j', t$	0.391***	0.574**	0.330***	0.315**
	(0.0826)	(0.220)	(0.107)	(0.143)
Constant	0.0144	-0.0228	0.0140	0.0652**
	(0.0157)	(0.0323)	(0.0281)	(0.0283)
Subject	All	Math	Reading	Math
Grade Level	All	Elementary	Elementary	Middle
Direction	A $\rightarrow B$	A $\rightarrow B$	A $\rightarrow B$	A $\rightarrow B$
Implied SD of School VA Change	.16	.19	.15	.14
Number of School Pairs	309	108	111	42

(b) Transitions from both School A to School B and from School B to School A

	(1) $\Delta r_j, t$	(2) $\Delta r_j, t$	(3) $\Delta r_j, t$	(4) $\Delta r_j, t$
$\Delta r_j', t$	-0.305***	-0.531***	-0.319*	-0.172
	(0.101)	(0.187)	(0.157)	(0.148)
Constant	-0.0187	-0.0731	-0.0823	0.110**
	(0.0254)	(0.0502)	(0.0550)	(0.0497)
Subject	All	Math	Reading	Math
Grade Level	All	Elementary	Elementary	Middle
Direction	B $\rightarrow A$	B $\rightarrow A$	B $\rightarrow A$	B $\rightarrow A$
Implied SD of School VA Change	.14	.19	.14	.11
Number of School Pairs	120	30	34	26

Notes: This table reports estimates of the relationship between changes in average student residuals for pairs of teachers who move between the same pair of schools, yielding a regression analogue to the raw variation visualized in figure 2. For each teacher j, we compute the change in average student residuals at the time of the move, Δr_{jt}. We then estimate regressions of the form $\Delta r_{jt} = \beta_1 \Delta r_{j't'} + \varepsilon_{j,t}$, where (j, j') index pairs of teachers moving between the same origin-destination schools. Panel (a) restricts attention to teacher pairs making the same move (e.g., from school s to s'), while Panel (b) considers teacher pairs making opposite moves (one from s to s' and the other from s' to s). Under the presence of school effects, we expect $\beta_1 > 0$ for same-move pairs and $\beta_1 < 0$ for opposite-move pairs. Columns denote subsamples corresponding to different subject-by-grade level combinations. The row labeled "Implied SD of School VA Change" reports $\sqrt{Cov(\Delta r_{j,t}, \Delta r_{j',t'})}$ which is equal to the standard deviation of changes in school Value-add ($\sqrt{Var(\Delta \mu_s)}$) provided measurement error is uncorrelated. Standard errors are clustered at the level of the origin-destination school pair with statistical significance denoted as follows: $***p < 1\%, **p < 5\%, *p < 10\%$.

interpretation depends in large part on the specific nature of teacher transitions. For instance, if teacher transitions are completely random, then $Var(\Delta \mu_s) = 2 \cdot Var(\mu_s)$. On the other hand, if teachers only move between schools with very similar value-added, then

$Var(\Delta\mu_s) \approx 0$ regardless of the true variance of school effects. Thus, in the next section, we descriptively characterize the network of school movers. The findings from this descriptive analysis of the network will motivate a more formal identification argument, which we develop in section VII.

VI Understanding Teacher Transitions

As discussed in the preceding section, the network of teacher moves is key to disentangling teacher and school effects. However, the appropriate way to leverage this variation depends critically on the nature of the moves. For that reason, we now provide a descriptive characterization of movers and the corresponding network created by those moves. This analysis yields three important facts that will motivate our identification argument in section VII.

Our first key finding is that teachers who move are observably different than teachers who stay. Table 2 presents results from a t-test comparing means of demographic characteristics for movers and stayers. Movers are significantly more likely to be black and female, and significantly less likely to be white. They also have 1.8 fewer years of experience than stayers on average. Movers' students also score 0.19 standard deviations higher on ELA tests but score similar to stayers' students on Math tests. Note that for movers, this statistic averages across years before and after moving.

Our second key finding is that the transitions themselves are not random. Appendix Table 5 describes differences in the students taught by a mover before and after the move. Movers tend to transition to classes that are more advantaged. For example, post-move classrooms tend to have a higher share of academically gifted students, a lower share of students that have a disability diagnosis on file, and a lower share of students that are classified as economically disadvantaged. Teacher moves are also associated with a 0.1 SD increase in both Math and ELA test scores–perhaps, unsurprising given the nature of the other observable changes. Test score gains associated with moves are of similar magnitude for each move (e.g. first move, versus second move, versus third), although the small number

Table 2: Characteristics of Movers vs Stayers

Teacher Characteristics	Stayers	Movers	Difference
Ethnicity: White	0.804	0.775	0.029
			(0.006)***
Ethnicity: Black	0.152	0.188	-0.036
			(0.005)***
Ethnicity: Hispanic	0.022	0.017	0.005
			(0.002)***
Gender: female	0.884	0.896	-0.012
			(0.004)***
Age	39.965	38.303	1.662
			(0.143)***
Years of Experience	11.543	9.759	1.784
			(0.109)***
Teacher's Highest Educational Qualification:			
Doctoral Degree	0.004	0.004	-0.000
			(0.001)
Sixth Year (Advanced) Degree	0.006	0.005	0.001
			(0.001)
Master's Degree	0.353	0.363	-0.009
			(0.007)
Bachelor's Degree	0.637	0.629	0.009
			(0.007)
Student Test Score: Math	-0.172	-0.169	-0.003
			(0.011)
Student Test Score: ELA	-0.226	-0.189	-0.037
			(0.011)***
Number of Teachers	35600	6162	
Avg Years Observed	2.576	3.788	

Notes: This table compares teachers who move between schools at least once during the study period ("Movers") to those who do not ("Stayers"). Columns (1) and (2) report group means, and Column (3) reports differences in means (Stayers minus Movers), with standard errors in parentheses from two-sample t-tests allowing for unequal variances. Teacher characteristics are averaged across all years a teacher is observed. Student test scores are standardized within subject, grade, and year, and reported as the average achievement of a teacher's students across all observed years; for movers, these averages include both pre- and post-move periods. $***p < 0.01$, $**p < 0.05$, $*p < 0.10$.

of third-time movers leaves us under-powered to detect significant changes in those cases.

Figure 3 provides a helpful, visual summary of the findings discussed in the preceding paragraph. To summarize the factors that explain classroom level academic gains after a move, we decompose observed test scores into two components–a predicted score component, which is the part of the variation that can be explained by the student observables in Table 5, and a residual component that contains the remaining variation.[15] Notably, the remaining variation could be coming from the teacher's quality, the quality of the match between the teacher and their students, or any other unobserved factors. We then compare test scores for movers and stayers before and after their first move in an event study framework.[16] The first move provides the most precise estimates since there are a much large number of first movers ($N = 6,162$) as compared to second or third moves ($N = 686$ and $N = 34$, respectively). Gains in the predicted score component almost entirely explain gains in the observed test scores for ELA, and explain nearly three-fourths of the increase for Math. This suggests that changes in student composition are an important factor leading to the observed test score gains that tend to happen after teachers transition.

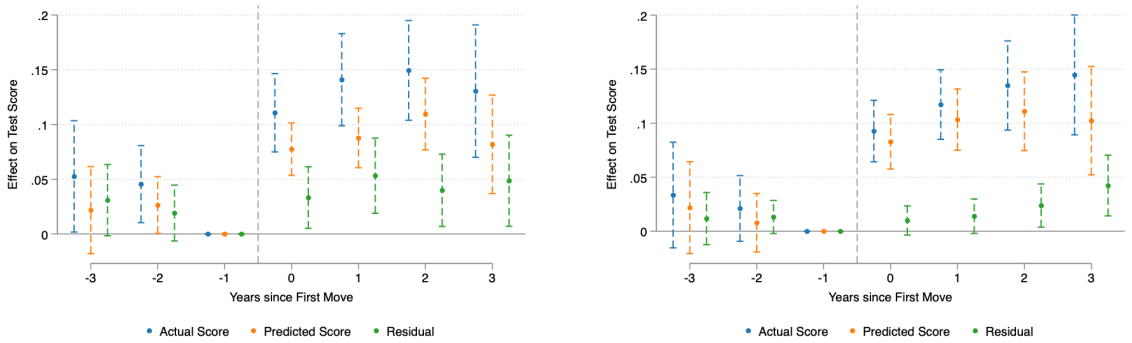
Our third and final descriptive finding is that the network of teacher transitions is sparse. This is because there are relatively few moves overall. Figure 4 displays the network visually. Within it, each node is a school. Edges denote an origin destination pair connected by at least one move. Edges highlighted in red represent weakly connected sub-networks–those that are highly leveraged (in the sense of Kline et al. (2020)). Figure 5 displays a histogram of these highly leveraged points.[17] As is clear from the figure, many of the leverage values equal 1, which implies that the assumptions necessary to estimate the variance components

[15]In practice, we form the predictive component by estimating a linear model of the form $y_{i,t} = \beta X_{i,t} + \epsilon_{i,t}$. $y_{i,t}$ are test scores in year t for student i, and $X_{i,t}$ are student covariates including lagged test scores, gender, race/ethnicity, and indicators for their LEP status, disability status, whether they are economically disadvantaged or academically gifted. We then calculate the predicted value ($\hat{y}_{i,t}$) from this regression. This is the predicted score, and the residual is the error term ($\epsilon_{i,t}$).

[16]Formally, we average the actual scores, predicted scores and residuals from our predictive regression at the teacher-year level. We then estimate a descriptive, event-study type model of the form $y_{j,t} = \sum_{k \neq -1} \beta_k D_{j,k} + \phi_j + \kappa_t + \epsilon_{j,t}$. $D_{j,k}$ takes value 1 for movers starting in the year of their first move ($k = 0$). ϕ_j and κ_t are teacher and year fixed effects respectively.

[17]Leverage here is defined based on the regression in Section VII.A.

Figure 3: Event Study: Teacher's First Move



(a) Math

(b) ELA

This figure plots estimates of event-study coefficients β_k estimated from teacher-year level regressions of the form $y_{j,t} = \sum_{k \neq -1} \beta_k D_{j,k} + \phi_j + \kappa_t + \epsilon_{j,t}$, where $D_{j,k}$ indexes event time relative to a teacher's first move ($k = 0$), and ϕ_j and κ_t denote teacher and year fixed effects, respectively. $y_{j,t}$ are teacher-year averages of student scores, predicted scores, and residuals constructed from regressing test scores on student covariates. The regression specification used is $y_{i,t} = \beta X_{i,t} + \epsilon_{i,t}$. See more details in Section VI.

using jack-knife methods are violated (see Lemma 1 in Kline et al. (2020)). Thus, the structure of weak network ties in our data make it difficult to estimate the sampling variances necessary to distinguish signal from noise. This motivates several important features of the identification argument that we will develop in the next section.

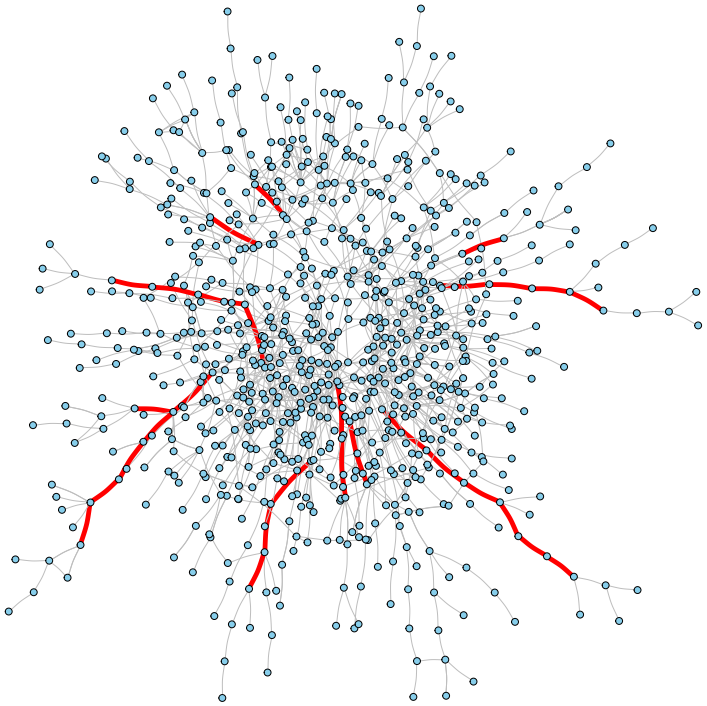
VII Econometric Model and Estimation Approach

We now aim to turn the suggestive evidence in Section V and the teacher transitions described above to understand the importance of teachers and schools on student test scores. We start by formally specifying the econometric model and assumptions before outlining our approach and then (in the rest of the paper) describing the results.

VII.A Generating Unbiased School Effect Estimates

We start by outlining how we can use the network of teacher transitions to construct estimates of school effectiveness and valid measures of their precision. In doing so there are two main challenges: (a) that teacher transitions may be endogenous; and (b) that linkages

Figure 4: The Network of Teacher Transitions



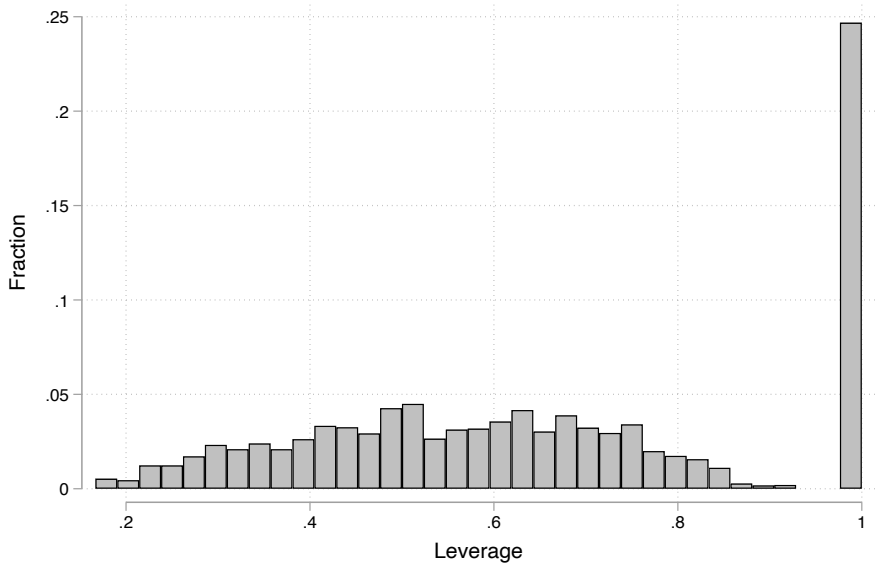
Notes: This figure highlights the largest network of connected teacher transitions in our data (the "superconnected set" in the language of Card et al. (2013)). Each node is a school. Edges denote an origin destination pair connected by at least one move. Edges highlighted in red represent weakly connected sub-networks–those with the highest leverage. The weakly connected sub-networks make it difficult to estimate the sampling variances necessary to distinguish signal from noise, and motivate several features of our identification argument.

among schools are often weak, i.e., a single teacher moved between them.[18]

To handle the potential endogeneity of teacher transitions, we introduce an identification approach that does not rely on comparing teachers who switch schools ("movers") to teachers who do not ("stayers"), as is the case in traditional AKM models (Abowd et al.

[18]There is a third challenge, in that not not all schools can be compared using teacher transitions, i.e., there are multiple connected components in the network of teacher transitions. As is common, we ignore this complication and focus on schools within the largest connected component.

Figure 5: Leverage Histogram



Notes: This is a histogram of each observations' leverage, based on the regression discussed in Section VII.A pooling over both subjects and grade-levels. Almost one-quarter of the observations have a leverage of one, suggesting that removing those observations from the dataset would render some parameters unable to be estimated. In practice, this usually corresponds to schools that are connected to the larger connected component by a single teacher transition.

(1999)), and instead relies on comparing movers to other movers. To see how this works, consider the stylized example in Figure 6. In the top panel, Figure 6a, we see a district of six schools (labeled as nodes A through F) with the teacher transitions illustrated via lines. For example, a teacher moved from D to A and from A to B, but not from A to D or D to B.

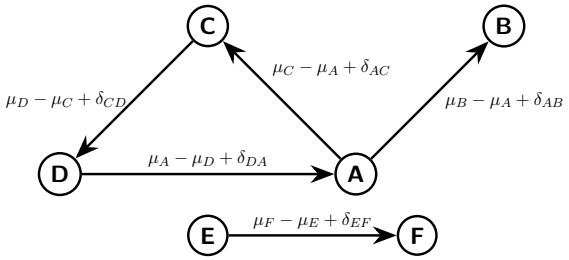
For each transition, we can calculate the change in the teachers' residual. Consider, for example, if teacher j moved from school A in year $t-1$ to school B in year t; using Equation 7 we get that:

$$r_{jt} - r_{jt-1} = \big(\mu_B - \mu_A\big) + \big((\theta_{jt} + \epsilon_{jt}) - (\theta_{jt-1} + \epsilon_{jt-1})\big) \tag{10}$$

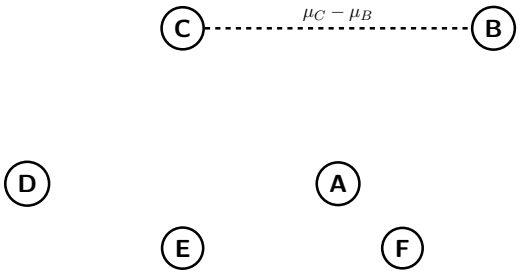
In other words, teacher j's observed residuals may change between the two years due

Figure 6: Identification Intuition

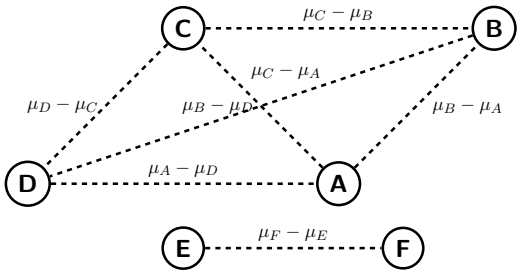
(a) Network of Transitions



(b) Assume $\delta_{ij} = \delta_i$ for all $i \neq j$



(c) Assume $\delta_{ij} = \delta$ for all $i \neq j$



to differences in SVA between school A and school B (i.e., $\mu_B - \mu_A$). Alternatively, this change may instead be because teacher j's TVA changed (i.e., $\theta_{jt} - \theta_{jt-1}$), either because School A is a better or worse fit for teacher j than School B or because , or because of year-over-year changes in the error (i.e., non-zero $\epsilon_{jt} - \epsilon_{jt-1}$, perhaps due to changes in unobservable student characteristics).

Consider then not the realized change, but the expected change in the teachers' measured residual. Treating the SVA measures as fixed parameters, we then get that: $\mathbb{E}[r_{jt} - r_{jt-1} | j \in J_{At-1} \cap J_{Bt}] = \mu_A - \mu_B + \delta_{ij}$ where:

$$\delta_{AB} = \mathbb{E}\Big[\big(\theta_{jt} + \epsilon_{jt}\big) - \big(\theta_{jt-1} + \epsilon_{jt-1}\big) \big| j \in J_{At-1} \cap J_{Bt}\Big]$$

As is clear from Figure 6a, we cannot identify the differences in school quality without making some assumptions on the δ_{ij} terms. One possibility is to assume that $\delta_{ij} = \delta_i$ for all $i \neq j$, i.e., that the two teachers who move from the same school to different schools with the same SVA would (in expectation) see the same change in their measured residual. While

this assumption permits some identification, it requires many transitions between many schools to identify the differences in school effects. This is illustrated in Figure 6c, which shows that under this assumption only the difference between μ_C and μ_B is identified.[19] We will therefore generally consider a stronger version of this assumption, that $\delta_{ij} = \delta$ for all $i \neq j$. Roughly, this implies that two teachers who switch between schools with similar differences in SVA would (in expectation) see the same change in their measured residual. Using $m_{j,t}$ to indicate whether teacher j switched schools between year $t-1$ and year t, this can be specified as:

Assumption 1. *Limited Transition Endogeneity*

$$\begin{aligned} &\mathbb{E}\Big[\big(\theta_{j,t} + \epsilon_{j,t}\big) - \big(\theta_{j,t-1} + \epsilon_{j,t-1}\big) \big| j \in J_{st} \cap J_{s't-1}\Big] = \\ &\mathbb{E}\Big[\big(\theta_{j,t} + \epsilon_{j,t}\big) - \big(\theta_{j,t-1} + \epsilon_{j,t-1}\big) \big| m_{j,t}\Big] \end{aligned} \tag{11}$$

We note that this assumption allows for many of the natural concerns one might have with using teacher transitions to identify school effects. For example, it allows for the process of moving to be disruptive and to temporarily cause a reduction in teacher effectiveness (i.e., $\theta_{j,t} - \theta_{j,t-1}$ being lower for movers than for stayers) or, alternatively, for teachers to generally move from schools where they are a poor fit to schools where they are a better fit (i.e., $\theta_{j,t} - \theta_{j,t-1}$ being higher for movers than for stayers). It also does not rule out the possibility that teachers who had a particularly difficult set of students (i.e., a negative value of $\epsilon_{j,t}$) are more likely to switch schools, even if this is true in ways unobserved by the researcher, or that principals assign teachers a particularly easy (i.e., high $\epsilon_{j,t}$) or difficult (i.e., low $\epsilon_{j,t}$) class after they switch schools. Furthermore, the assumption allows for teachers' effectiveness to evolve over time rather than assuming that it is a static latent variable and makes few assumptions on how it does so.

The key assumption is that these different explanations do not systematically depend on which schools they move between. In other words, it rules out the possibility that teachers

[19] Note that there are actually no teacher transitions between these two schools, and the only way they are "connected" is because teachers from School A moved to both.

moving from School A to School C do so in search of better match effects, while teachers moving from School D to School A do so because they were teaching a particularly difficult set of students. To be clear, it is not an issue if some teachers move in search of better match effects, while others do so after they were teaching a particularly difficult set of students; instead, the assumption states that we cannot infer why a specific teacher moved based on the schools she moved between.

As shown in Figure 6c, this assumption allows us to identify the differences between school value-added of all the schools in each connected component of the graph. To see how this assumption permits identification, suppose that the teachers who move from School A to School C consistently see an increase in their average residual. Unlike traditional models of employee/employer matching, we do not assume that this implies that School C is necessarily better than School A, since it may instead be due to the reasons outlined in the paragraph above. If we also see that teachers who move from School A to School B do not consistently see an increase in their average residual and instead generally see a decrease in their average residual, we can – if not rule out those explanations – at least know that they work similarly in the two instances. This means that by comparing the how teachers change when they move from School A to School C to how they change when the move from School A to School B, we can identify the difference between school value-added in School B to school value-added in School C. This can be done using a difference-in-difference style estimator, where the first difference – how teachers change when the move from School A to School B or School C – gives estimates of $\mu_B - \mu_A + \delta_{AB}$ and $\mu_C - \mu_A + \delta_{AC}$ and the second difference – between those two transitions – gives estimates of $\mu_B - \mu_C$ since, by assumption, $\delta_{AB} - \delta_{AC} = 0$.

More generally, we can use the cycle of transitions where teachers move from School A to School C, School C to School D, and from School D to School A. Averaging the change in teachers' residuals from these three changes must give δ, since the average change in school quality must net to zero (given that it's a cycle). With the knowledge of δ, we can then identify differences in school quality between any schools connected by a series of teacher

transitions.

This identification logic suggests the following regression:

$$\Delta r_{jt} = \mu'\big(D_{jt} - D_{jt-1}\big) + \delta' m_{jt} + e_{jt} \tag{12}$$

where D_{jt} is a vector of binary variables indicating which school teacher j taught in year t.[20] Indeed, from Assumption 1 it immediately follows that $\mathbb{E}[\hat{\mu}] = \mu$, where $\hat{\mu}$ are the estimated regression coefficients and μ is the vector of SVA.[21] This also highlights how we can relax Assumption 1 by allowing for the different sources of endogeneity to differ depending, for example, whether a teacher moves within-district or across-districts or whether a teacher is relatively young or old. To do so, all we need to do is to include different measures m_{jt} depending on the type of move and/or characteristic of the teachers and allow for the δ term to consist of multiple parameters.

While useful, the relatively small number of teachers switching schools means that each estimate $\hat{\mu}$ contains substantial uncertainty. A particular challenge in this context is that even among schools that are linked it is often the case that only a single teacher moves between them. An important consequence of this is that we perfectly predict many of the observations in the regression that we use, which implies that we cannot simply estimate the uncertainty in these estimates using the squared residuals to approximate the variance of the error terms. We can see this empirically in Figure 5, which shows that leverage of nearly one-quarter of the observations are equal to one after running regressions of the form Equation 12. This implies that the only way to generate variance estimates is to add more structure; see Theorem 1 of Kline et al. (2020). We therefore add the following assumption:

Assumption 2. *Limited Heteroskedasticity*

$$\big(\theta_{jt} + \epsilon_{jt}\big) - \big(\theta_{jt-1} + \epsilon_{jt-1}\big) | m_{jt} = 1 \sim N\big(\delta, \sigma^2_{jt}\big) \tag{13}$$

[20] When running the regression we need to normalize the SVA measures somehow, which we do by including a sum-to-zero constraint. This can be done in the regression by include an additional observation with $\Delta r_{jt} = 0$ and where $D_{jt} = 1$ for all schools and $m_{jt} = 0$.

[21] Assumption 1 implies that $\mathbb{E}[e_{jt}|D_{jt}, D_{jt-1}, m_{jt}] = \mathbb{E}[e_{jt}|m_{jt}] = 0$, which implies that result.

where:

$$\sigma_{jt}^2 = \beta' Z_{jt} + \xi_{jt} \tag{14}$$

for i.i.d. mean-zero error ξ_{jt} with finite variance.

The crucial assumptions here are twofold. First, embedded in this formulation is the assumption that the error is independent across transitions. In fact, when not focused on a single subject (e.g., looking across both math and English estimates) we allow the error to be correlated across subjects within the same teacher; we discuss that in Appendix A, but ignore it here for notional simplicity. Second, we permit limited forms of heteroskedasticity by allowing the variance of each observation to depend on a number of teacher characteristics (Z_{jt}). This allows us to use portions of the network where the school effects can be estimated using a leave-one-out approach to understand the variance of the error and then extrapolate using these estimates to portions of the network where the structure breaks down using a leave-one-out approach. Finally, we also assume that the error term is distributed normally; however, this is only necessary when conducting inference on the variance estimates and can be relaxed using asymptotic arguments.

We next outline our approach to estimating unbiased school effect estimates as follows:

Estimating School Value-Added. *We estimate school value-added according to the following steps:*

1. *Construct estimates of μ via a regression of the form in Equation 12.*

2. *For those observations with leverage (in the above regression) less than one, compute:*

$$\hat{e}_{jt}^2 = \Delta r_{jt} \cdot \left(\Delta r_{jt} - \hat{\mu}_{-j}(D_{jt} - D_{jt-1}) - \hat{\delta}_{-j} m_{jt} \right) \tag{15}$$

   where $\hat{\delta}_{-j}$ and $\hat{\mu}_{-j}$ denote leave-j-out estimates of the coefficients.

3. *Using these observations to estimate $\hat{\beta}$ by running a regression of the form:*

$$\hat{e}_{jt}^2 = \beta' Z_{jt} + u_{jt} \tag{16}$$

4. *Use the estimates of $\hat{\beta}$ to calculate $\hat{\sigma}^2_{jt} = \hat{\beta}' Z_{jt}$ for every observation (including those with leverage of one). Let $\hat{\Sigma}$ be the error matrix, with $\hat{\sigma}^2_{jt}$ on the diagonals and zeros everywhere else.*

5. *We can then estimate the variance of μ as:*

$$\hat{\Sigma}_{\hat{\mu}} = (D'D)^{-1} D' \hat{\Sigma} D (D'D)^{-1} \tag{17}$$

where D is the coefficient matrix from Equation 12.

We can then justify this approach with the following remark, which highlights that the above approach gives unbiased estimates of both the individual SVA estimates and the overall variance of SVA:

Remark 1. *If Assumptions 1 and 2 hold and $\hat{\mu}$ and $\hat{\Sigma}_{\hat{\mu}}$ are estimated according to the definition above, we have that:*

$$\hat{\mu} \sim N\big(\mu, \Sigma_{\mu}\big) \tag{18}$$

$$\mathbb{E}\big[\hat{\Sigma}_{\mu}\big] = \Sigma_{\mu} \tag{19}$$

VII.B Constructing Unbiased Variance and Covariance Estimates

We next turn our attention to understanding the overall importance of teachers and schools on student test scores, as well as an estimate of how teachers sort across schools based on teacher effectiveness. For this, we will use the unbiased estimations of SVA to construct multiple variance and covariance estimates of interest.

To do so, we start by making an assumption common to nearly all value-added approaches, namely that the error is independent across time.[22] To formalize this, we start

[22] As discussed in more detail below, the persistence of the bias in teacher value-added that either ignore or try to adjust for school value-added differences is what distinguishes this type of error from general misspecification of the value-added estimates.

with the specification (discussed above) that:

$$r_{jt} = \theta_{jt} + \mu_s + \epsilon_{jt} \tag{20}$$

Assumption 3. *We assume that the error term is distributed as follows:*

$$\epsilon_{jt} \sim N\big(0, \sigma^2_\epsilon\big) \tag{21}$$

Note that this assumes both that the error is independent across teachers and across time within teachers.[23] Note that the error term combines both the classroom shock (i.e., the error common to all students in the class) and the average of the individual error terms (i.e., shocks that hit each student individually). For simplicity, here we assume that the variance of this is common across teachers (e.g., σ^2_ϵ); in Appendix A, we separate the error into individual and common error to account for the fact that teachers have different numbers of students. We will find it useful to define $\overline{r}_s$ as the average value of r_{jt} for every j, t in school s, i.e., $\overline{r}_s = |N_S|^{-1} \sum r_{jt}$ if there are N_S number of teacher-year observations in school s. Similarly, we will define $\overline{\theta}_s$ as $|N_S|^{-1} \sum \theta_{jt}$ and $\overline{\epsilon}_s$ as $|N_S|^{-1} \sum \epsilon_{jt}$.

Given the low fraction of movers we will make one more simplifying assumption, that the error from estimating school effectiveness is uncorrelated with the general error in r_{jt}. Formally, the assumption is as follows:

Assumption 4. *Let ν_s be the error in the school value-added estimates, i.e., $\hat{\mu}_s = \mu_s + \nu_s$. Then*

$$Cov(\epsilon_{jt}, \nu_s) = 0 \tag{22}$$

$$Cov(\overline{\epsilon}_s, \nu_s) = 0 \tag{23}$$

Note error in $\mu_s - \hat{\mu}_s$ stems only from characteristics of the movers, while the overall

[23] As discussed in Appendix A, we allow for correlation across subjects within a specific teacher in a given year.

covariances are (almost) completely calculated using characteristics of the stayers, which suggests that the assumption that the error terms are uncorrelated is reasonable.[24]

We are primarily interested in estimating the following variances and covariances:

1. $\sigma_s^2 \equiv V(\mu_s)$, which reflects how much schools vary in their effectiveness;
2. $\sigma_{\overline{\theta}}^2 \equiv Var(\overline{\theta}_s)$, which reflects how much schools vary in the average effectiveness of the teachers;
3. $\sigma_{\mu,\overline{\theta}} \equiv Cov(\mu_s, \overline{\theta}_s)$ and $\sigma_{\mu,\theta} \equiv Cov(\mu_s, \theta_{jt})$, which reflect how related school quality and teacher quality at the schools are;
4. $\sigma_{\theta,|t-t'|} \equiv Cov(\theta_{j,t}, \theta_{j,t'})$, which reflects how much teachers vary in their effectiveness, accounting for potential drift between years t and t'.

Supposing that we know σ_ϵ^2, we can then estimate the true variances using the observed moments as follows:

$$\hat{\sigma}_\mu^2 = Var(\hat{\mu}_s) - trace(\hat{\Sigma}_\mu) \tag{24}$$

$$\hat{\sigma}_{\overline{\theta}}^2 = Var(\overline{r}) - \hat{\sigma}_\mu^2 - 2 \cdot \hat{\sigma}_{\mu,\overline{\theta}} - \frac{1}{N_S}\sigma_\epsilon^2 \tag{25}$$

$$\hat{\sigma}_{\mu,\overline{\theta}} = Cov(\hat{\mu}_s, \overline{r}_s) - \hat{\sigma}_\mu^2 \tag{26}$$

$$\hat{\sigma}_{\mu,\theta} = Cov(\hat{\mu}_s, r_{jt}) - \hat{\sigma}_\mu^2 \tag{27}$$

$$\hat{\sigma}_{\theta,|t-t'|} = Cov(r_{jt} - \hat{\mu}_s, r_{jt'} - \hat{\mu}_s') - trace(S_t \hat{\Sigma}_\mu S_{t'}) \tag{28}$$

where S_t corresponds to a matrix with a value of one in element i, j if individual i attended school j in year t and zero otherwise.

The approach gives us unbiased estimates of the true variance measures, as stated in the following remark:

Remark 2. *If Assumptions 2, 3, and 4 hold and the estimation approach is as outlined*

[24] Furthermore, it depends on year-to-year changes in the error term, rather than the levels.

above, we have that:

$$\mathbb{E}\big[\hat{\sigma}^2_{\mu}\big] = \sigma^2_{\mu} \tag{29}$$

$$\mathbb{E}[\hat{\sigma}^2_{\overline{\theta}}] = \sigma^2_{\overline{\theta}} \tag{30}$$

$$\mathbb{E}\big[\hat{\sigma}_{\mu,\overline{\theta}}\big] = \sigma_{\mu,\overline{\theta}} \tag{31}$$

$$\mathbb{E}\big[\hat{\sigma}_{\mu,\theta}\big] = \sigma_{\mu,\theta} \tag{32}$$

$$\mathbb{E}\big[\hat{\sigma}_{\theta,|t-t'|}\big] = \sigma_{\theta,|t-t'|} \tag{33}$$

We leave the proof of this to Appendix A, where we also show how we can construct inference on these variance estimates. We also discuss there how we can estimate σ^2_{ϵ} and $\sigma^2_{\theta} \equiv Var(\theta_{j,t})$.

VII.C Empirical Bayes Estimates of Teacher and School Value-Added

Although much can be learned from the variance and covariance estimates, much of the research on teacher and school value-added is aimed at constructing individual estimates of teacher or school quality to be used by policy makers, principals, and/or evaluators. We therefore now turn to the question of how we can best use the available data to construct such estimates.

To do so, we employ a two-step empirical Bayes (EB) approach, in which we first combine the unbiased (but noisy) measures of school quality (the $\hat{\mu}$ estimates) with precise (but biased) measures of school quality (the $\overline{r}_s$ values) to construct posterior SVA measures. We then use these posterior SVA measures to adjust the teacher-level residuals and finally use these adjusted teacher-level residuals to construct posterior TVA measures.

For the first step, we start with the following prior on the joint distribution of SVA and average TVA at each school:

$$\begin{bmatrix} \mu_s \\ \overline{\theta}_s \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{\mu} & \sigma_{\mu,\overline{\theta}} \\ \sigma_{\mu,\overline{\theta}} & \sigma^2_{\overline{\theta}} \end{bmatrix} \right) \tag{34}$$

where the components of the variance of the prior are terms that we have estimated in the prior section. In practice, we will plug in our estimates of the variance components rather than their true values. Using the previously assumed structure of the error term and the assumed prior, we can then derive:

$$\overline{r}_s|\mu_s \sim N\Big(\mu_s \cdot \big(1 + \frac{\sigma_{\mu,\overline{\theta}}}{\sigma^2_\mu}\big), \sigma^2_{\overline{\theta}} - \frac{\sigma^2_{\mu,\overline{\theta}}}{\sigma^2_\mu} + \sigma^2_\epsilon \cdot \frac{1}{N_S}\Big) \tag{35}$$

$$\hat{\mu}|\mu \sim N\big(\mu, \Sigma_\mu\big) \tag{36}$$

Letting S_μ be a diagonal matrix where the diagonals are equal to σ^2_μ, $\tilde{S}_{\overline{r}}$ is a diagonal matrix with the diagonals equal to $\big(\sigma^2_{\overline{\theta}} - \frac{\sigma^2_{\mu,\overline{\theta}}}{\sigma^2_\mu} + \frac{\sigma^2_\epsilon}{N_S}\big)$, and $\alpha = 1 + \frac{\sigma_{\mu,\overline{\theta}}}{\sigma^2_\mu}$, and $S^{-1}_{\overline{r}} = \alpha^2 \cdot \tilde{S}^{-1}_{\overline{r}}$, we can then use Bayes Law to invert the distribution and get that:

$$\mu|\hat{\mu}, \overline{r} \sim N\Big(\hat{\mu}^{EB}, \Sigma_{SVA,post}\Big) \tag{37}$$

where $\Sigma_{SVA,post} = \big(S^{-1}_\mu + \Sigma^{-1}_\mu + S^{-1}_{\overline{r}}\big)^{-1}$ and $\hat{\mu}^{EB} = \Sigma_{post} \cdot \big(\Sigma^{-1}_\mu \hat{\mu} + S^{-1}_{\overline{r}} \alpha^{-1} \overline{r}\big)$.

Given these posterior estimates of school quality, we can then move on to construct posterior estimates of teacher quality. For this, we start by adjusting the teacher-level residuals by subtracting from them our empirical Bayes estimate of school quality. Formally, we will let $\tilde{r}_{jt} = r_{jt} - \hat{\mu}^{VA}_s$, from which if follows that:

$$\tilde{r}_{jt} = \theta_{jt} + \big(\mu_s - \hat{\mu}^{EB}_s\big) + \epsilon_{jt}$$

This equation highlights that there are now two reasons why the adjusted residual does not fully represent teacher quality: first, the traditional i.i.d. shock corresponding to ϵ_{jt}; second, the fact that the empirical Bayes estimates of school quality ($\hat{\mu}^{EB}_s$) do not necessarily correspond to the true school quality (μ_s). Because this source of error is not independent across years, it is important that we separate the two sources of error. If we let $\nu_s = \mu_s - \hat{\mu}^{EB}_s$, we can not that even though ν_s is not independent over time, we know that $\nu_s \sim N(0, \sigma^2_{s,EB})$ where $\sigma^2_{s,EB}$ comes directly from the posterior distribution we derived

above.

We simplify by assuming that there are enough teachers in the school to ensure that ν_s and ϵ_{jt} are independent. We can then use standard results in empirical Bayes methods to get that:

$$\mathbb{E}[\theta_{jt}|\tilde{r}_{jt}] = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\epsilon^2 + \sigma_{s,SVA}^2} \cdot \tilde{r}_{jt}$$

VIII How Much do Teachers and Schools Matter for Student Test Scores?

We start with the variance estimates, which shed light on the relative importance of teachers and schools in determining student test scores and help us understand how much teachers sort based on value-added across North Carolina.

The main results are shown in Table 3 below. A number of important conclusions can be drawn from these results. First, as seen in the first row, we can conclude that schools are indeed more than a collection of teachers. In particular, when pooling over both subjects and grade-levels, we find that the estimated standard deviation of SVA (σ_μ) is 0.10 and can reject the null that there is no variation in SVA.[25]

For comparison, the estimated standard deviation of TVA is 0.13, and so is slightly higher than the estimated standard deviation of SVA. In other words, although school-wide policies can indeed affect student test scores, having a one-standard deviation above average teacher does more for a student test scores than attending a one-standard deviation above average school.

The last result highlights that teachers do not significantly sort across schools based on their TVA. In particular, the standard deviation in average TVA at each school ($\sigma_{\bar{\theta}}$) is less than half as large as the standard deviation of individual TVA. We also find (not shown in the Table 3) that there is no correlation between the average TVA at each school and the

[25] Our method provides unbiased estimates of the variance of SVA, but for ease of interpretation we report the standard deviations of SVA. Note that the non-linear transformation of an unbiased parameter means that the reported standard deviations estimate are likely biased.

SVA. In other words, there is no evidence that better teachers consistently sort to better schools.[26]

Table 3: Estimated Variation in Teacher and School Quality

Sample	(1) Full	(2) Math	(3) Reading	(4) Elementary	(5) Middle
σ_μ	0.10	0.11	0.08	0.10	0.10
	(0.07,0.12)	(0.07,0.14)	(0.02,0.11)	(0.05,0.13)	(0.06,0.12)
σ_θ	0.13	0.17	0.08	0.13	0.12
	(0.11,0.14)	(0.15,0.19)	(0.02,0.10)	(0.11,0.16)	(0.10,0.14))
$\sigma_{\bar{\theta}}$	0.06	0.08	0	0.06	0.04
	[0,0.08)	(0.03,0.11)	[0,0.07)	[0,0.10)	[0,0.08)
Number of Schools	2,315	1,107	1,208	1,434	881
Number of Teachers	65,871	31,066	34,805	39,001	26,870

Notes: This table reports the estimated standard deviations of school value-added (σ_μ), TVA (σ_θ), and avg. TVA at the school level ($\sigma_{\bar{\theta}}$). We use the method outlined in Section VII and Appendix A to construct variance estimates and 95% confidence intervals on these estimates. In the table above, we report the estimated standard deviations and CIs by taking the square root of the variance estimate and the range of the 95% CIs. If the 95% CI range includes a negative values, we truncate the range at 0 and indicate that it is inclusive of zero.

The next four columns of Table 3 show that these results are generally consistent across both subjects and across both grade-levels. Comparing Column (2) to Column (3), we find that both schools and teachers matter more for student math scores than student reading scores. This is a common result in the teacher and school value-added literature. In contrast, as shown in Column (4) and Column (5) that the variance effects are quite similar for elementary schools as they are for middle schools.

IX Implications for Theory and Practice

Having established our identification results and jointly estimated teacher and school effects from the data, we now demonstrate their practical utility (1) for our theoretical understanding of educational production; and (2) for informing the practical, day-to-day decisions that

[26]Formally, we cannot reject the null that the covariance of avg. TVA and SVA is zero.

district and school administrators must make. However, before we describe the specific exercises / thought experiments that justify these conclusions, it is useful to briefly discuss the broad classes of questions / policy problems where our method is likely to yield different results.

IX.A Where are existing school value-added methods insufficient?

In general, there are two use-cases where researchers and policy makers need worry less about separately identifying teacher and school effects. The first use case is when the question of interest depends only on a within-school ranking of teachers. For example, suppose a school principle must fire one teacher (without replacement) due to budget constraints. If they wish to maximize student test scores, it is sufficient to rank teachers within school and (up to uncertainty due to sampling variation) release the bottom performer.

However, many policy questions do not take this form and instead require a global ranking. For example, and as discussed in the introduction, many teacher evaluation systems require teacher ratings to be based on rankings that implicitly use cross-school variation. As another example, teacher vacancies are frequently filled through within-district transfers, reflecting institutional rules that give priority to incumbent teachers (e.g., Schweizer, 1999). From the principals perspective, estimating the impact of replacing a low-performing teacher with another teacher within-district therefore fundamentally requires a ranking across schools. This motivates the policy thought-experiment we will explore in section IX.C, which involves the frequency of type-1 errors and the magnitude of the expected losses when using traditional value-add for personnel decisions of this nature.

The second use case concerns questions where the combined influence of teachers and schools is a sufficient statistic. For example, a parent shopping for the best learning environment for their child may not care if a school is good because the school is effective or if a school is good because the teacher labor force is effective—they simply care that the combined influence of both teachers and schools delivers the largest benefit for their child, which is precisely the estimand of a traditional school value-added model.

However, many questions cannot be answered with a combined parameter and, instead, *require* a joint distribution. For example, there is a large literature on teacher mobility and sorting across schools (e.g., Hanushek et al., 2004; Boyd et al., 2005; Clotfelter et al., 2006; Boyd et al., 2011; Bruhn et al., 2022). It is impossible to measure whether highly effective teachers tend to work in highly effective schools without a method of separately identifying the two concepts. More generally, teacher mobility in the presence of strong school effects can complicate the interpretation of even seemingly simple descriptive patterns in the data (e.g. the returns to experience), a point we turn to in the next section.

Finally, we note here that our method has the key advantage that virtually *any* question that can be reliably answered using a traditional value-add measure can also be answered using separately identified teacher and school value-add. Since our procedure characterizes a joint distribution, it is trivial to then construct a within school distribution (as in use case 1) or a combined parameter (as in use case 2). However, the converse is not true—without a procedure like the one we have built in this paper, it is not possible to reconstruct the joint distribution from traditional value-added estimates. Thus, we would argue that there is little cost and (as we will demonstrate in the next two sections) much potential gain from estimating these concepts jointly, rather than separately.

IX.B Implications for Theory: The Returns to Teacher Experience

We begin with a classic question in the theory of educational production: how does teacher effectiveness evolve with experience? A large empirical literature finds substantial improvements early in teachers' careers, followed by more gradual gains thereafter (e.g., Rockoff, 2004; Papay and Kraft, 2015; Wiswall, 2013; Harris and Sass, 2011; Clotfelter et al., 2007). At the same time, it is well documented that teachers tend to move to schools with more advantaged student populations and higher observed achievement, particularly early in their careers (e.g. Boyd et al., 2011). To the extent that such schools also exhibit higher value-added, this mobility can induce a spurious correlation between experience and measured effectiveness.

To formalize this idea, consider the value-add regression in Rockoff (2004), which is representative of standard approaches in the literature:

$$r_{jt} = \beta \, \text{Exp}_{jt} + \alpha_j + X'_{it}\gamma + \varepsilon_{it}, \tag{38}$$

where r_{jt} denotes the teacher-level results estimated in Section **??**, Exp_{jt} is teacher experience, and α_j is a teacher fixed effect.

Now suppose the existence of an independent school effect. In that case, the error term can be decomposed as

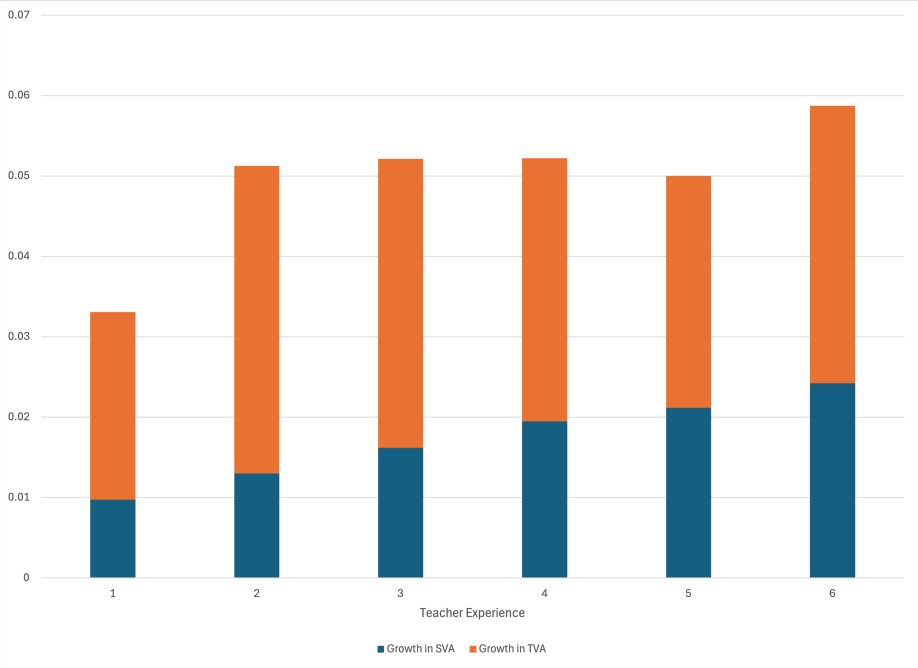
$$\varepsilon_{it} = \mu_{s(j,t)} + u_{it}, \tag{39}$$

where $\mu_{s(j,t)}$ captures the value-added of the school to which teacher j is assigned in period t. This raises the specter of omitted variable bias which, in this case, is given by: $\frac{Cov(\text{Exp}_{jt}, \mu_{s(j,t)})}{Var(\text{Exp}_{jt})}$. Thus, if teachers systematically move to higher value-added schools as they gain experience (e.g. $Cov(\text{Exp}_{jt}, \mu_{s(j,t)}) > 0$), then this would imply that the estimated return to experience overstates the true effect. Intuitively, even with teacher fixed effects, improvements in school quality that coincide with career progression are mechanically attributed to experience.

Motivated by this possibility, we use our jointly identified school and teacher value-added estimates to decompose the returns to experience into components that can be explained as a result of teachers moving to better schools, and a component that reflects actual returns to experience within-teacher over time. Figure 7 displays the results. We construct this figure by first estimating a Rockoff-style experience profile using the student-level regression $r_{it} = \alpha_j + \sum_k \beta_k \mathbf{1}\{\text{Exp}_{jt} = k\} + \varepsilon_{it}$, where r_{jt} is the teacher residuals, α_j is a teacher fixed effect, and $\mathbf{1}\{\text{Exp}_{jt} = k\}$ is an indicator for teacher j having k years of experience in year t. The height of each full bar gives the naive estimated return to experience at each experience level. The blue portion of each bar shows the component that disappears once we purge school value-added from the left-hand-side outcome variable by replacing r_{jt} with

$Y_{jt} - \hat{\mu}_{s(j,t)}$, where $\hat{\mu}_{s(j,t)}$ is our estimate of the value-added of the school to which teacher j is assigned in year t.

Figure 7: Decomposing Returns to Experience into Teacher and School Components



Notes: This figure decomposes estimated returns to teacher experience into a component attributable to teachers moving to higher value-added schools and a component reflecting within-teacher improvement over time. To do so, we first estimate a Rockoff-style experience profile in our data using the student-level regression $r_{jt} = \alpha_j + \sum_k \beta_k \mathbf{1}\{\text{Exp}_{jt} = k\} + \varepsilon_{it}$, where r_{jt} is the teacher-level residual, α_j is a teacher fixed effect, and $\mathbf{1}\{\text{Exp}_{jt} = k\}$ is an indicator for teacher j having k years of experience in year t. The height of each full bar gives the naive estimated return to experience at each experience level. The blue portion of each bar shows the component that disappears once we purge school value-added from the left-hand-side outcome variable by replacing Y_{it} with $Y_{it} - \hat{\mu}_{s(j,t)}$, where $\hat{\mu}_{s(j,t)}$ is our estimate of the value-added of the school to which teacher j is assigned in year t.

We find that a substantial portion of the apparent gains to experience can be explained as a result of teachers sorting to more effective schools. As is clear from figure 7, anywhere from 30-40% of the returns to experience is explained by teacher moving to schools with higher value-added, an fact that would be impossible to detect using standard value-added methods alone.

IX.C Implications for Practice: Identifying Effective Teachers

Next we turn to implications for personnel policy. Our thought experiment considers a principle who must decide between hiring a transfer teacher, and a random draw from the teacher labor pool, in order to fill an opening. As discussed earlier, personnel problems of this form, where the decision depends on the global distribution (in this case the global average teacher versus a transfer), not just the within school distribution, are commonly faced by principles and districts. Importantly, because the principle is considering a transfer it is reasonable to assume they have access to growth scores or some other form of value-added estimate that they may use to inform their decision. However, we do not assume the principal is a test-score maximizer— just that they wish to have an accurate forecast of how their decision will impact test scores as part of the policy process.

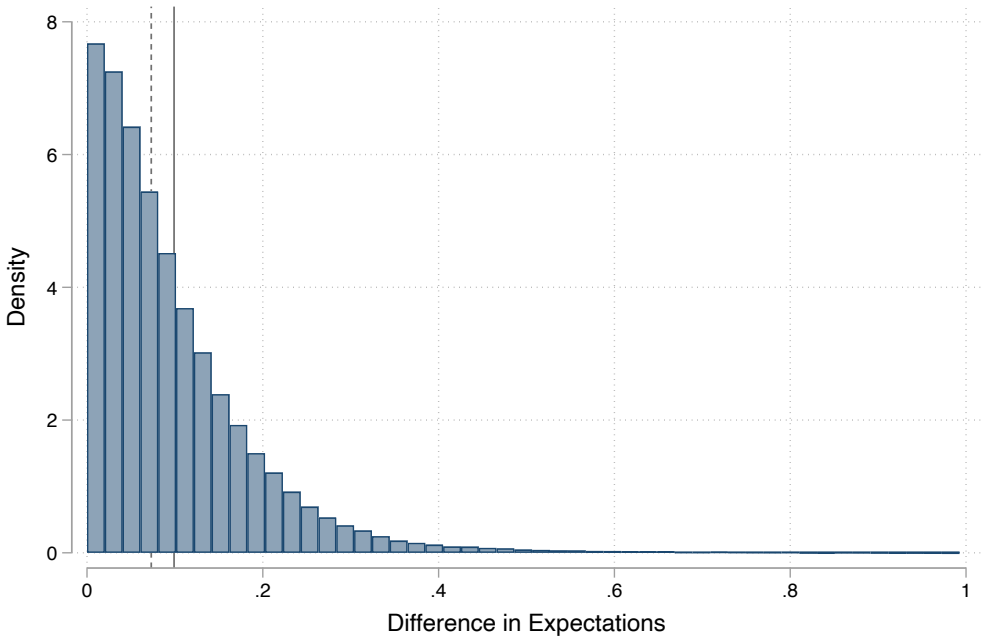
More formally, suppose the principle has access to a teacher value-added measure that does not separately identify schools, which we will denote by $\hat{\theta}^{VA}$. In this case we ask, "What is the gain relative to expectation if they decline to hire the transfer teacher and instead accept an average draw from the teacher labor pool?"[27] Mathematically, this forecast boils down to calculating, for each teacher: $\Delta\theta = |\mathbb{E}(\theta_{jt}|r_jst, \hat{\mu}) - \hat{\theta}^{VA}|$ where $\hat{\theta}_j^{VA}$ denotes the conventional value-added estimate that combines teacher and school components, r_{jst} denotes the underlying signal used to recover teacher effectiveness, and $\hat{\mu}$ denotes our separately identified estimate of school value-added. The object $\mathbb{E}[\theta_j \mid r_{jst}, \hat{\mu}]$ is the posterior expectation of teacher j's true teacher-specific value-added after separately accounting for the contribution of the school. Note that we take the absolute value, since we are interested in the *gains relative to expectation*, not the overall distribution of both improvements and losses, since this is a natural measure of how "wrong" the principal will be in terms of the expected change in value-add should they choose to decline the transfer.

Figure 8 plots the histogram of gains relative to expectation. The horizontal axis reports $\Delta\theta_j$, which measures the magnitude of the expected improvement in teacher value-added

[27] In practice, similar results would obtain if we specified specific, alternative teachers. In that way, benchmarking to a hypothetical average teacher is just a normalization

from making the correct personnel decision for teacher j, relative to replacing that teacher on the basis of the conventional measure $\hat{\theta}_j^{VA}$. The solid vertical line marks the mean of the distribution, and the dashed vertical line marks the median.

Figure 8: Distribution of Gains Relative to Expectation (Standard TVA)



Notes: This figure plots the distribution of gains relative to expectation from replacing a teacher when personnel decisions are based on a value-added measure that does not separately identify school effects. For each teacher j, we define the expected gain as $\Delta\theta_j = \left|\mathbb{E}[\theta_j \mid r_{jst}, \hat{\mu}] - \hat{\theta}_j^{VA}\right|$, where $\hat{\theta}_j^{VA}$ denotes the conventional value-added estimate that combines teacher and school components, r_{jst} denotes the underlying signal used to recover teacher effectiveness, and $\hat{\mu}$ denotes our separately identified estimate of school value-added. The object $\mathbb{E}[\theta_j \mid r_{jst}, \hat{\mu}]$ is the posterior expectation of teacher j's true teacher-specific value-added after separately accounting for the contribution of the school. The horizontal axis reports $\Delta\theta_j$, which measures the magnitude of the expected improvement in teacher value-added from making the correct personnel decision for teacher j, relative to replacing that teacher on the basis of the conventional measure $\hat{\theta}_j^{VA}$. We take the absolute value because the object of interest is the size of the gain from correcting the decision, rather than the signed difference itself. The solid vertical line marks the mean of the distribution, and the dashed vertical line marks the median.

As is evident from figure 8, the principals forecast of the test score impacts will, on average, be off by a substantial margin. Specifically, the average error implies a change in actual test scores that is roughly half as large as expected. This occurs because the principal is making their forecast using value-add that includes the independent school effect; since

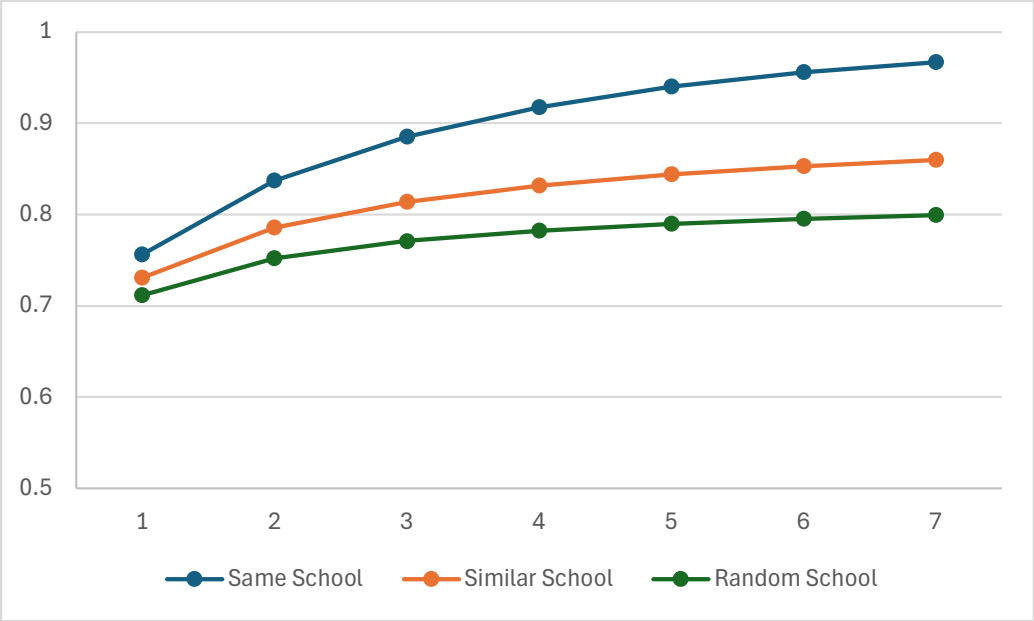
these school effects represent a non-trivial share of the variance (see section **??**), they create systematic and large deviations in the expected impact of the policy. This is consistent with evidence on the *actual* causal effect of adopting value-added based assessment systems on student outcomes, which finds treatment effects that are substantially smaller than what could otherwise theoretically be obtained (Gutierrez et al., 2016; Stecher et al., 2018; Bleiberg et al., 2025).

Another way to see this is to note that, within the context of a standard value-added estimator, the independent school component functions as a source of measurement error that will not disappear asymptotically, even with infinite data. This implies that personnel decision will retain sensitivity to the reference school regardless of the length of time a teacher has spent in district or the number of students they have taught. However, if the variance of the school component is small relative to the variance of the sampling error (as might theoretically be the case for experienced teachers who have taught many students), then it could be the case that addressing biases from school effects is second order relative to improvements in precision—a topic important enough to personnel policy that has been the subject of an exhaustive literature (see Koedel and Rockoff, 2015a, for a review).

We explore this possibility empirically in Figure 9. The figure answers a simple question: Suppose $\theta_{jt} - \theta_{j't} = \sigma_\theta$; that is we wish to compare two teachers where the first is a full standard deviation better (in a value-added sense) than the second. How likely is it that $\hat{\theta}^{VA}_{jt} > \hat{\theta}^{VA}_{j't}$, and how does magnitude of the type–1 error rate vary with additional years of teaching experience (which is the primary way the effective student N can increase to improve precision for an individual teacher)? In the figure, the x-axis denotes years of experience. The Y-axis gives the type-1 error rate. The three series correspond to three different reference groups: a comparison teacher drawn from the same school; a comparison teacher drawn from a school that is similar in terms of school quality;[28] and a random school.

[28]This is formalized by restricting comparisons to teachers at schools within one-standard deviation of SVA.

Figure 9: Even in Large Samples, Personnel Decisions with Standard TVA are Sensitive to the Reference Group



Notes: This figure plots the likelihood that a conventional teacher value-added estimator correctly ranks teachers when one of them is a full standard deviation better than the other. The three series correspond to drawing the reference teacher from the same school as teacher j, from a school that is within one-standard deviation in quality, or from a random school in the district. Under a standard TVA estimator, $\hat{\theta}^{VA}_{j,t} = \theta_{j,t} + \mu_{s(j,t)} + e_{j,t}$, where $e_{j,t}$ vanishes as T grows but the school component $\mu_{s(j)}$ does not. As a result, same-school comparisons become nearly exact in large samples because the school term cancels, while cross-school comparisons remain sensitive to the reference group even as $T \to \infty$.

As is clear from figure 9, school effects are a first order issue for identifying more effective teachers. With few years of experience (e.g. 1) we observe the correct rates of identification are high (in part because we have assumed a large difference in value-add between the two hypothetical teachers), but that the type-1 error rate is similar across all three school environments (≈0.3-0.24). This is due to the fact that, early on in a teacher's career, there is little data by which to form this likelihood (often as low as 12-20 students in a single classroom) and hence the noise dominates the signal. However, as the number of years of experience grows, we find a sharp diverge in the ability of a standard value-added measure to correctly identify the better teacher *within*-school versus *across*-school. Because the within school comparison implicitly nets out the independent school component, we find that by year 7, after a substantial accumulation of data, type-1 error rates become low (≈0.05), whereas they remain high for a random school (≈0.2) and even those that are similar on observables (≈0.13). Importantly, this exercise is done under the (conservative) assumption of a 1-SD difference in value-added between the two teachers; our results would become even more start if the difference were small. In that way, this exercise demonstrates the substantive importance of accounting for independent school effects for teacher personnel policy.

X Conclusion

This paper asks whether schools are more than a collection of teachers. Using administrative data from North Carolina and the network of teacher transitions across schools, we develop a framework for jointly identifying teacher and school value-added and show that this distinction is economically important. Our evidence indicates that schools exert independent effects on student achievement that conventional value-added models confound with teacher quality, and that this confounding has first-order implications for both research and policy. In particular, we find that standard methods overstate the returns to teacher experience because teachers move to better schools early in their careers, and that across-

school personnel decisions based on conventional teacher value-added remain error-prone even with very large samples because the school component does not vanish asymptotically. More broadly, our results suggest that evaluation systems built on a single combined measure can obscure central features of educational production and misdirect accountability and staffing decisions. Joint estimation therefore does more than refine existing measures: it changes the objects that researchers and policymakers can recover, making it possible to study teacher quality, school effectiveness, and the allocation of educators across schools within a unified framework.

References

Abdulkadiroglu, Atila, Parag A. Pathak, and Christopher R. Walters, "Who Gets What in Education: Can School Matching Improve Student Achievement?," Technical Report 34936, National Bureau of Economic Research 2026.

Abowd, John M., Francis Kramarz, and David N. Margolis, "High Wage Workers and High Wage Firms," *Econometrica*, 1999, *67* (2), 251–333.

Andrews, Martyn J., Len Gill, Thorsten Schank, and Richard Upward, "High Wage Workers and Low Wage Firms: Negative Assortative Matching or Limited Mobility Bias?," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 2008, *171* (3), 673–697.

Bartanen, Brendan and Aliza N. Husain, "Connected Networks in Principal Value-Added Models," *Economics of Education Review*, 2022, *90*, 102292.

Bates, Michael, Michael Dinerstein, Andrew C. Johnston, and Isaac Sorkin, "Teacher Labor Market Policy and the Theory of the Second Best," *Quarterly Journal of Economics*, 2025.

Bleiberg, Joshua, Eric Brunner, Erica Harbatkin, Matthew A. Kraft, and Matthew G. Springer, "Taking Teacher Evaluation to Scale: The Effect of State Reforms on Achievement and Attainment," *Journal of Political Economy Microeconomics*, 2025, *3* (3), 568–610.

Bonhomme, Stéphane, Thibaut Lamadon, and Elena Manresa, "A Distributional Framework for Matched Employer Employee Data," *Econometrica*, 2019, *87* (3), 699–739.

Boyd, Donald, Hamilton Lankford, Susanna Loeb, and James Wyckoff, "Explaining the Short Careers of High-Achieving Teachers in Schools with Low-Performing Students," *American Economic Review Papers & Proceedings*, 2005, *95* (2), 166–171.

_ , _ , _ , **Matthew Ronfeldt, and James Wyckoff**, "The Role of Teacher Quality in Retention and Hiring: Using Applications to Transfer to Uncover Preferences of Teachers and Schools," *Journal of Policy Analysis and Management*, 2011, *30* (1), 88–110.

Bruhn, Jesse, Scott Imberman, and Marcus Winters, "Regulatory arbitrage in teacher hiring and retention: Evidence from Massachusetts charter schools," *Journal of Public Economics*, 2022, *215*, 104750.

Card, David, Jörg Heining, and Patrick Kline, "Workplace Heterogeneity and the Rise of West German Wage Inequality," *Quarterly Journal of Economics*, 2013, *128* (3), 967–1015.

Cheng, Xu, Sheng Chao Ho, and Frank Schorfheide, "Optimal Estimation of Two-Way Effects under Limited Mobility," NBER Working Paper 34014, National Bureau of Economic Research 2025.

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

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

_ , _ , **and** _ , "Teacher Credentials and Student Achievement: Longitudinal Analysis with Student Fixed Effects," *Economics of Education Review*, 2007, *26* (6), 673–682.

Cortes, Kalena, Karen Kortecamp, Susanna Loeb, and Carly Robinson, "A Scalable Approach to High-Impact Tutoring for Young Readers: Results of a Randomized Controlled Trial," Technical Report, National Bureau of Economic Research 2024.

Covelli, Lauren, John Engberg, and Isaac M. Opper, "Leading Indicators of Long-Term Success in Community Schools: Evidence from New York City," 2024.

Dee, Thomas S. and James Wyckoff, "Incentives, Selection, and Teacher Performance: Evidence from IMPACT," *Journal of Policy Analysis and Management*, 2015, *34* (2), 267–297.

District of Columbia Public Schools, "IMPACT: The District of Columbia Public Schools Effectiveness Assessment System for School-Based Personnel," Technical Report, District of Columbia Public Schools 2010. Technical report describing the IMPACT teacher evaluation system.

Eastmond, Tanner S., Michael Ricks, Julian Betts, and Nathan Mather, "Welfare Added? Optimal Teacher Assignment with Value-Added Measures," Technical Report 34768, National Bureau of Economic Research 2026.

Florida Department of Education, "Florida VAM Methodology," Technical Report, Florida Department of Education 2015. Incorporated in Rule 6A-5.0411, Calculations of Student Learning Growth for Use in School Personnel Evaluations, Effective August 2015.

Gilraine, Michael and Nolan G. Pope, "Making Teaching Last: Long-Run Value-Added," NBER Working Paper 29555, National Bureau of Economic Research December 2021.

_ , Jiaying Gu, and Robert McMillan, "A New Method for Estimating Teacher Value-Added," NBER Working Paper 27094, National Bureau of Economic Research May 2020.

Gutierrez, Italo A., Gabriel Weinberger, and John Engberg, "Improving Teaching Effectiveness: Impact on Student Outcomes: The Intensive Partnerships for Effective Teaching Through 2013–2014," Technical Report RR-1295/3-1-BMGF, RAND Corporation, Santa Monica, CA 2016.

Hanushek, Eric A., John F. Kain, and Steven G. Rivkin, "Why Public Schools Lose Teachers," *Journal of Human Resources*, 2004, *39* (2), 326–354.

Harris, Douglas N. and Tim R. Sass, "Teacher Training, Teacher Quality, and Student Achievement," *Journal of Public Economics*, 2011, *95* (7-8), 798–812.

Jackson, C Kirabo, "Match quality, worker productivity, and worker mobility: Direct evidence from teachers," *Review of Economics and Statistics*, 2013, *95* (4), 1096–1116.

Jochmans, Koen and Martin Weidner, "Fixed-Effect Regressions on Network Data," *Econometrica*, 2019, *87* (5), 1543–1560.

Kane, Thomas J. and Douglas O. Staiger, "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation," Technical Report w14607, National Bureau of Economic Research 2008.

Kane, Thomas J, Daniel F McCaffrey, Trey Miller, and Douglas O Staiger, "Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment. Research Paper. MET Project.," *Bill & Melinda Gates Foundation*, 2013.

Kinsler, Josh, "Assessing Rothstein's Critique of Teacher Value-Added Models," *Quantitative Economics*, 2012, *3* (2), 333–362.

Kline, Patrick, Raffaele Saggio, and Mikkel Solvsten, "Leave-Out Estimation of Variance Components," *Econometrica*, 2020, *88* (5), 1859–1898.

Koedel, Cory and Jonah E Rockoff, "Value-added modeling: A review," *Economics of Education Review*, 2015, *47*, 180–195.

_ **and Jonah E. Rockoff**, "Value-Added Models in Education: A Primer for Policymakers," *Journal of Economic Perspectives*, 2015, *29* (3), 93–112.

_ **and Julian R. Betts**, "Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique," *Education Finance and Policy*, 2011, *6* (1), 18–42.

Kraft, Matthew A., John P. Papay, Jessalynn K. James, and Manuel Monti-Nussbaum, "Is Teacher Effectiveness Fully Portable? Evidence from the Random Assignment of Transfer Incentives," NBER Working Paper 34845, National Bureau of Economic Research 2026.

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

Mulhern, Christine, "Beyond Teachers: Estimating Individual School Counselors' Effects on Educational Attainment," *American Economic Review*, 2023, *113* (11), 2846–2893.

_ **and Isaac Opper**, "Measuring and Summarizing the Multiple Dimensions of Teacher Effectiveness," 2021.

Papay, John P. and Matthew A. Kraft, "Productivity Returns to Experience in the Teacher Labor Market," *Journal of Public Economics*, 2015, *130*, 105–119.

Rockoff, Jonah E., "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data," *American Economic Review Papers & Proceedings*, 2004, *94* (2), 247–252.

Rose, Evan K., Jonathan T. Schellenberg, and Yotam Shem-Tov, "The Effects of Teacher Quality on Adult Criminal Justice Contact," NBER Working Paper 30274, National Bureau of Economic Research July 2022. Revision dated September 2022.

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

Schweizer, Peter, "The Dance of the Lemons," January 1999. Accessed: 2026-03-23.

Stecher, Brian M., Deborah J. Holtzman, Michael S. Garet, Laura S. Hamilton, John Engberg, Elizabeth D. Steiner, Abby Robyn, Matthew D. Baird, Italo A. Gutierrez, Evan D. Peet, Iliana Brodziak de los Reyes, Kaitlin Fronberg, Gabriel Weinberger, Gerald P. Hunter, and Jay Chambers, "Improving Teaching

Effectiveness: Final Report: The Intensive Partnerships for Effective Teaching Through 2015–2016," Technical Report RR-2242-BMGF, RAND Corporation, Santa Monica, CA 2018.

Swisher, Abigail and Patricia Saenz-Armstrong, "National Council on Teacher Quality. State of the States 2022: Teacher and Principal Evaluation Policies," 2022.

Tartova, Desislava, "Teacher Value-Added in the Absence of Annual Test Scores: Utilizing Teacher Networks," 2023. SSRN working paper; revised February 2025.

Turque, Bill, "Creative, motivating — and fired," *The Washington Post*, 2012.

U.S. Department of Education, National Center for Education Statistics, "Common Core of Data (CCD): Public Elementary/Secondary School Universe Survey, Local Education Agency Universe Survey, and State Nonfiscal Survey of Public Elementary/Secondary Education, 2022–23, Provisional Version 1a," Dataset 2023. Public Elementary/Secondary School Universe Survey (2022–23, Provisional Version 1a); Local Education Agency Universe Survey (2022–23, Provisional Version 1a); State Nonfiscal Survey of Public Elementary/Secondary Education (2022–23, Provisional Version 1a).

Wiswall, Matthew, "The Dynamics of Teacher Quality," *Econometrica*, 2013, *81* (3), 1153–1191.

Woodcock, Simon D., "Match Effects," *Research in Economics*, 2015, *69* (1), 100–121.

A Econometric Details of Variance Estimates

In the main body of the paper, we outlined our estimation approach. Here, we provide additional details that highlight: 1) how we handle within-teacher/year across-subject correlations; 2) how we use a parametric model to separately identify σ^2_θ and from σ^2_θ; and 3) how to estimate the variance of the parameter estimates.

A.A Extended Model, Updated Notation, and Target Parameters

We will start by adding some additional notation, namely using matrix notation for most of the estimates. We will let $\hat{\mu}$ be the vector of coefficient estimates from Equation 12; note that this is a slight abuse of notation, as it includes not only the school effects but also the estimated δ (or other coefficients) from the regression. We write that $\hat{\mu} = \mu + \nu$, where ν is the vector of the error terms and let ΔR be the vector of $\Delta r_{j,t}$ values used to estimate this and D to be the matrix of covariates, so we can write: $\hat{\mu} = (D'D)^{-1}D'\Delta R$.

As in the main body of the paper, we will let Σ be the variance matrix corresponding to the errors of ΔR. We highlight how Assumption 2 can be extended to allow for within-teacher-year correlations of the error term.

Assumption 5. *The ΔR term is distributed as follows:*

$$\Delta R \sim N\big(\Delta\mu + \delta, \Sigma\big) \tag{40}$$

where Σ is a matrix with $Z\beta_{var} + \xi_{var}$ on the diagonals, $Z\beta_{cov} + \xi_{cov}$ on the ij terms that reflect the same teacher who is teaching different subjects in the same year, and zero elsewhere. Here, ξ_{var} and ξ_{cov} are i.i.d. error terms, with mean zero and finite variance.

As before, we estimate $\hat{\beta}_{var}$ as $(Z'I_pZ)^{-1}Z'I_pE$, where I_p is a diagonal matrix with the diagonal equal to one if the leverage, i.e., P_{jt}, of that observation is less than one and zero otherwise and E is a vector that corresponds to $\Delta r_{jt} \cdot (\Delta r_{jt} - d'_{jt}\hat{\mu}_{-jt})$, where $_{-jt}$ is a leave-jt-out estimate of $\hat{\mu}$. Similarly, we will estimate $\hat{\beta}_{cov}$ as $(Z'I_{pc}Z)^{-1}Z'I_{pc}C$

where I_{pc} is a diagonal matrix with the diagonal equal to one if the leverage, i.e., P_{jt}, of that observation is less than one and the subject is math and the teacher is an elementary school teacher (and so also has an ELA measure for that year), and zero otherwise. Here $C = \Delta r_{j't} \cdot (\Delta r_{jt} - d'_{jt}\hat{\mu}_{-jt})$ if teacher j and j' are the same individual teacher teaching different subjects. We will assume that Z is independent of the error term, as stated below:

Assumption 6. *We assume that* $\Delta R - D\mu \perp\!\!\!\perp Z$.

Note that this is a relatively benign assumption, as it can be achieved by including the coefficients than comprise of Z in the initial regression.

Similarly, we will let R be the vector of values of $r_{j,t}$ and S to be the matrix that maps individual teachers to schools; we can then write $R = \theta + S\mu + \epsilon$, where ϵ is the vector of error terms. We will write $\mathbb{E}[\epsilon \cdot \epsilon'] = \Sigma_\epsilon$. Again, we will extend Assumption 3 to be as follows:

Assumption 7. *We have that* Σ_ϵ *is a positive definite symmetric matrix with* $\sigma^2_{class} + \frac{1}{N_{jt}} \cdot \sigma^2_{sudent}$ *on the diagonal and zeros everywhere else, except for the* ij *terms that correspond to the same teacher in the same year (corresponding to different subjects). Here,* σ^2_{class} *is a shock common to all students in the class,* $\sigma^2_{student}$ *is the error term unique to each student, and* N_{jt} *is the number of students taught by teacher* j *in year* t.

When averaged at the school level, we will write $\overline{R}$, so again we get that: $\overline{R} = \overline{\theta} + \mu + \overline{\epsilon}$. Again, we will let $\Sigma_{\overline{\epsilon}} = \mathbb{E}[\overline{\epsilon} \cdot \overline{\epsilon}']$.

Motivated by the relatively small number of teachers who switch schools and the fact, we will also keep the assumption that the different error terms are independent. Formally, this assumption is listed below:

Assumption 8. *We have that* $\nu \perp\!\!\!\perp \epsilon$

Given this notation, we can then write the main target parameters in terms of the vectors as follows:

- Variance of school effects: $V(\mu_s) = \mu' A_1 \mu$

- Variance of average teacher effectiveness: $V(\overline{\theta}_s) = \overline{\theta}' A_1 \overline{\theta}$
- Amount of sorting on average teacher quality: $Cov(\mu_s, \overline{\theta}_s) = \mu' A_1 \overline{\theta}$
- Amount of sorting on individual teacher quality: $Cov(\mu_s, \theta_{jt}) = S'\mu' A_2 \theta$
- Variance of teacher effects: $Var(\theta_{jt}) = \theta' A_2 \theta$
- Persistence of teacher effects: $Cov(\theta_{jt}, \theta_{jt-k}) = \theta' A_{k+2} \theta$

where S is a matrix that maps teachers to schools and A_k are matrices that determine the weight on each observation, can select the sample (e.g., focusing on within-district variances or on middle school teachers), and/or shift the covariances within teachers over time (as in the case with A_4 and A_5 matrices).

A.B Estimation Approach and Unbiasedness

We start by defining the following moment conditions:

$$M = \begin{bmatrix} \hat{\mu}' A_1 \hat{\mu} - trace(A_1 \hat{\Sigma}_\mu) \\ \hat{\mu}' A_1 \overline{R} \\ \overline{R}' A_1 \overline{R} \\ \hat{\mu}' S' A_2 S \hat{\mu} - trace(S' A_2 S \hat{\Sigma}_\mu) \\ R' A_2 R \\ (R - S\hat{\mu})' A_2 (R - S\hat{\mu}) - trace(S' A_2 S \hat{\Sigma}_\mu) \\ (R - S\hat{\mu})' A_3 (R - S\hat{\mu}) - trace(S' A_3 S \hat{\Sigma}_\mu) \\ (R - S\hat{\mu})' A_4 (R - S\hat{\mu}) - trace(S' A_4 S \hat{\Sigma}_\mu) \end{bmatrix} \tag{41}$$

Lemma 1. *For appropriately chosen matrices $B_{k,ii}$ and $B_{k,cov}$ and the moment vector M*

defined above, we get that:

$$\mathbb{E}\big[M\big] = \begin{bmatrix} \mu' A_1 \mu \\ \mu' A_1 \overline{\theta} + \mu' A_1 \mu \\ \overline{\theta}' A_1 \overline{\theta} + 2 \cdot \overline{\theta}' A_1 \mu + \mu' A_1 \mu + trace(A_1 \Sigma_{\overline{\epsilon}}) \\ \mu' S' A_2 S \mu \\ \theta' A_2 \theta + 2 \cdot \mu' S' A_2 \theta + \mu' S' A_2 S \mu + trace(A_2 \Sigma_{\epsilon}) \\ \theta' A_2 \theta + trace(A_2 \Sigma_{\epsilon}) \\ \theta' A_3 \theta \\ \theta' A_4 \theta \end{bmatrix} \tag{42}$$

Proof. Using the fact that:

$$\begin{aligned} R - S\hat{\mu} &= (\theta + S\mu + \epsilon) - (S\mu + S\nu) \\ &= \theta - S\nu + \epsilon \end{aligned} \tag{43}$$

and assuming the weighting matrices are symmetric, we can re-write these as functions of the true parameters and the error terms as follows:

$$\begin{aligned} \hat{\mu} A_1 \hat{\mu} &= (\mu + \nu)' A_1 (\mu + \nu) \\ &= \mu' A_1 \mu + 2 \cdot \nu' A_1 \mu + \nu' A_1 \nu \end{aligned} \tag{44}$$

$$\begin{aligned} \hat{\mu}' A_1 \overline{R} &= (\mu + \nu)' A_1 (\overline{\theta} + \mu + \overline{\epsilon}) \\ &= \mu' A_1 \overline{\theta} + \mu' A_1 \mu + \nu' A_1 (\overline{\theta} + \mu) + \mu' A_1 \overline{\epsilon} + \nu' A_1 \overline{\epsilon} \end{aligned} \tag{45}$$

$$\begin{aligned} \overline{R}' A_1 \overline{R} &= (\overline{\theta} + \mu + \overline{\epsilon})' A_1 (\overline{\theta} + \mu + \overline{\epsilon}) \\ &= \overline{\theta}' A_1 \overline{\theta} + 2 \cdot \overline{\theta}' A_1 \mu + \mu' A_1 \mu + 2 \cdot \overline{\epsilon}' A_1 (\overline{\theta} + \mu) + \overline{\epsilon}' A_1 \overline{\epsilon} \end{aligned} \tag{46}$$

$$\begin{aligned} S'\hat{\mu}'A_2R &= (S\mu + S\nu)'A_2(\theta + S\mu + \epsilon) \\ &= \mu'S'A_2\theta + \mu'S'A_2S\mu + \mu'SA_2\epsilon + \nu'S'A_2(\theta + S\mu) + \nu'SA_2\epsilon \end{aligned} \tag{47}$$

$$\begin{aligned} (R - S\hat{\mu})'A_{k+3}(R - S\hat{\mu}) &= (\theta - S\nu + \epsilon)'A_{k+3}(\theta - S\nu + \epsilon) \\ &= \theta'A_{k+3}\theta - 2\cdot\nu'S'A_{k+3}\theta + \nu'S'A_{k+3}S\nu + \epsilon'A_{k+3}\epsilon - 2\cdot\nu'S'A_{k+3}\epsilon \end{aligned} \tag{48}$$

Using the fact that the error terms ϵ and ν are independent, we can calculate the expected value of these as follows:

$$\mathbb{E}\big[\hat{\mu}'A_1\hat{\mu}\big] = \mu'A_1\mu + trace(A_1\Sigma_\mu) \tag{49}$$

$$\mathbb{E}\big[\hat{\mu}'A_1\overline{R}\big] = \mu'A_1\overline{\theta} + \mu'A_1\mu \tag{50}$$

$$\mathbb{E}\big[\overline{R}'A_1\overline{R}\big] = \overline{\theta}'A_1\overline{\theta} + 2\cdot\overline{\theta}'A_1\mu + \mu'A_1\mu + trace(A_1\Sigma_{\bar{\epsilon}}) \tag{51}$$

$$\mathbb{E}\big[S\hat{\mu}'A_2R\big] = \mu'S'A_2\theta + \mu'S'A_2S\mu \tag{52}$$

$$\mathbb{E}\big[(R - S\hat{\mu})'A_{k+3}(R - S\hat{\mu})\big] = \theta'A_{k+3}\theta + trace(S'A_{k+3}S\Sigma_\mu) + trace(A_{k+3}\Sigma_\epsilon) \tag{53}$$

From the assumption that A_4 and A_5 are designed to look at the covariances between $(R - S\hat{\mu})$ across different years and the only non-zero values of Σ_ϵ are the diagonals and terms that correspond to the same teacher in the same year, we get that $trace(A_{k+3}\Sigma_\epsilon) = 0$.

Next, note that we can use the cyclic property of the trace to get that:

$$\begin{aligned} trace(S'A_lS\Sigma_\mu) &= trace(S'A_lS(D'D)^{-1}D'\Sigma D(D'D)^{-1}) \\ &= trace(\underbrace{D(D'D)^{-1}S'A_lS(D'D)^{-1}D'}_{B_l}\Sigma) \end{aligned} \tag{54}$$

where Σ is the variance matrix of the error terms in the regression to estimate unbiased school value-added.[29] Furthermore, based on our assumptions about Σ we can write this

[29]The matrix B_1 is defined in the same way, without the S and S' terms.

as:

$$\begin{aligned} trace(B_l\Sigma) = B'_{l,ii}\sigma_i^2 + B'_{l,cov}\sigma_{ij}^2 \\ B'_{l,ii}(Z\beta_{var} + \xi_{var}) + B'_{l,cov}(Z\beta_{cov} + \xi_{cov}) \end{aligned} \tag{55}$$

where $B_{l,ii}$ is the diagonal of B_l and $\beta_{l,cov}$ is a vector based on B corresponding to select cases where the covariances are not equal to zero.

Finally, we note that:

$$\begin{aligned} \Delta r_{jt} \cdot (\Delta_{rj} - d'_{jt}\hat{\mu}_{-jt}) &= (d'_{jt}\mu + e_{jt}) \cdot (d'_{jt}(\mu - \hat{\mu}_{-jt}) + e_{jt}) \\ &= d'_{jt}\mu \cdot d'_{jt}(\mu - \hat{\mu}_{-jt}) + e_{jt} \cdot d'_{jt}(\mu - \hat{\mu}_{-jt}) + e_{jt} \cdot d'_{jt}\mu + e_{jt}^2 \end{aligned}$$

Using the fact that $\hat{\mu}_{-jt}$ is an unbiased estimate of μ and unrelated to e_{jt} and ξ_{var} is a mean-zero and independent error term, we get that $\mathbb{E}\big[\Delta r_{jt} \cdot (\Delta_{rj} - d'_{jt}\hat{\mu}_{-jt})|D, Z\big] = \mathbb{E}[e_{jt}^2|D, Z] = z'_{jt}\beta_{var}$. Thus, $\mathbb{E}[\hat{\beta}_{var}] = \beta_{var}$. We can prove that $\mathbb{E}[\hat{\beta}_{cov}] = \beta_{cov}$ in the same way. Thus, we get that $trace(A_k\hat{\Sigma}_\mu)$ is an unbiased estimate of $trace(A_k\Sigma_\mu)$ for any A_k.

□

This Lemma gets us close to the result, but is not quite sufficient. There are two main issues. First, expected value of the empirical moments is often not directly the estimand of interest; for example, the expected value of $\hat{\mu}'A_1\overline{R}$ is a linear combination of both $\mu'A_1\mu$ and $\mu'A_1\overline{\theta}$. This is easy to handle via a simple linear transformation of the proposed moments. The second is much more challenging and comes from the presence of $trace(A_3\Sigma_\epsilon)$ and $trace(A_1\Sigma_{\overline{\epsilon}})$ in the expectations. In fact, we cannot separately identify $trace(A_3\Sigma_\epsilon)$ and $\theta'A_3\theta$ – or similarly $\overline{\theta}'A_1\overline{\theta}$ and $trace(A_1\Sigma_{\overline{\epsilon}})$ – without more structure.

This lack of identification is a common feature of teacher value-added models and comes from the fact that it requires understanding how much of a teacher's overperformance in a given year stems from a year specific shock to the external factors such as there being a particularly (and unobservably) disruptive student in the class and how much stems from a year-specific shock to true teacher productivity such as the teacher missing multiple weeks

due to a temporary health issue such as a broken arm.

The TVA literature has handled this challenge in two primary ways. The first is to use the fact that some teachers teach multiple classrooms to try and separate these shocks. The second is to use a parametric model, including one where there is a persistent component of teacher effectiveness and an (ignorable) yearly shock. Given our lack of quality classroom data for most teachers and our short panel, use a linear model in which we assume that $Cov(\theta_{j,t}, \theta_{jt-k}) = \rho_1 + \rho_2 k$. Our results are robust to other approaches, however, including the bounding assumption that $\theta' A_3 \theta \approx \theta' A_4 \theta$. Finally, we will estimate the individual student error $\sigma^2_{student}$ using the across-student variance of the residuals after netting out teacher-by-year fixed effects. Given the large numbers of students, we will estimate this separately from the rest of the moments and consider this fixed for all inference estimates discussed in the next section.

Given all this, we can then estimate the variance parameters as $W^{-1}M$, where M is defined above and W is defined as:

$$W = \begin{bmatrix} 1,0,0,0,0,0,0,0,0 \\ 1,1,0,0,0,0,0,0,0 \\ 1,2,1,0,0,0,0,tr(A_1 N_c^{-1}) \\ 0,0,0,1,0,0,0,0 \\ 0,0,0,1,2,1,0,tr(A_2) \\ 0,0,0,0,0,1,0,tr(A_2) \\ 0,0,0,0,0,1,1,0 \\ 0,0,0,0,0,1,2,0 \end{bmatrix} \tag{56}$$

where N_c^{-1} is a diagonal matrix corresponding to the number of classes in each school. Note that $tr(A_2) = 1$, which further simplifies things.

$$P = \begin{bmatrix} \mu' A_1 \mu \\ \mu' A_1 \overline{\theta} \\ \overline{\theta}' A_1 \overline{\theta} \\ \mu' S' A_2 S \mu \\ \mu' S' A_2 \theta \\ \rho_1 \\ \rho_2 \\ \sigma^2_{class} \end{bmatrix} \tag{57}$$

Theorem 1. *Given the definitions of P, M, and W above and our estimates $\hat{P}$, we get that:*

$$\mathbb{E}[\hat{P}] = P \tag{58}$$

A.C Variance of the Estimated Parameters

To estimate the variance of these estimated parameters, we now make two additional simplifications, both of which we plan to relax in future versions. The first is that the uncertainty in the parameters is dominated by the teachers who are weakly connected to the main connected component. The second assumption we make further restricts the allowable heteroskedasticity by assuming that that the variance of Δr_{jt} is exactly equal to $\beta' Z_{jt}$ or, equivalently, that $\xi_{jt} = 0$ for all jt observations. In practice, this means in the vector M above we can treat the estimates $\hat{\mu}$ as being the only source of uncertainty. With these assumptions, we can use common results on the variance of linear combinations and quadratic forms to calculate the variance-covariance matrix of M, which we can then use in a straightforward manner to determine the variance-covariance matrix of $\hat{P}$.

B Sample Restrictions

Our sample restrictions follow the same general approach as Chetty, Friedman, and Rockoff (2014). Estimating teacher and school effects requires lagged test scores. Consequently, we cannot calculate these effects in the first year of observed data (2015), though we use 2015 data to estimate teacher effects in years 2016-2019. We omit students who are not matched to a teacher in the data, students with missing test scores, students and teachers who move schools within a year, and students who are assigned to multiple teachers for the same subject within a year. We also exclude 8th grade math test scores for 2018 and 2019 due to changes in North Carolina testing guidelines.[30]

Table 4: Relationship Between Teacher Residual Changes and Differences in School Average Residuals

	(1)	(2)	(3)	(4)	(5)
	$\Delta r_j, t$	$\Delta r_j, t$	$\Delta r_j, t$	$\Delta r_j, t$	$\Delta r_j, t$
$\Delta \mu^{full}_s, t$	0.786^{***}	0.732^{***}	0.877^{***}	0.709^{***}	0.913^{***}
	(0.0299)	(0.0466)	(0.0549)	(0.0635)	(0.0694)
Subject	All	Math	Reading	Reading	Math
Grade Level	All	Elementary	Elementary	Middle	Middle
School Transitions	5269	1771	1720	865	913

$***p < 1\%, **p < 5\%, *p < 10\%$.

C Appendix Tables

D Measuring Years of Experience

Estimating the returns to experience requires a precise measure of the number of years a teacher has been teaching. We construct a measure of a teacher's years of experience based

[30]Before 2018, all 8th grade students took the same end-of-grade math test. In 2018, advanced math students began taking a different test to regular math students. Consequently, post-2017 cohorts lack a common measure of math test performance across all students. As a result, we exclude 8th grade math test scores for 2018 and 2019 from the analysis. See ESSA Defined Exception for Advanced Mathematics in Grade 8: https://support.wyzant.com/hc/en-us/articles/360021799571-Terms-of-Use

Figure 10: Relationship Between Teacher Residual Changes and Differences in School Average Residuals

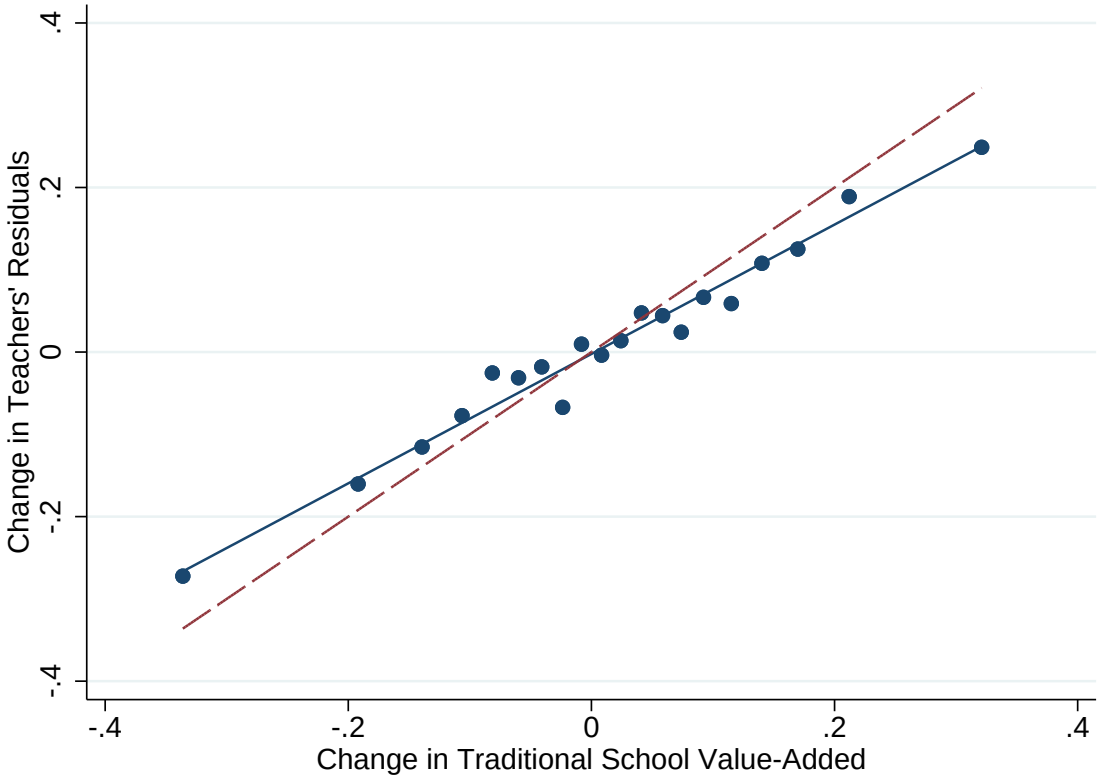


Table 5: Students Taught by Mover Before and After a Move

	Move 1	Move 2	Move 3
Panel A- Student Demographics			
Share of White students	0.038	0.039	-0.010
	(0.005)***	(0.016)**	(0.071)
Share of Black students	-0.045	-0.057	-0.028
	(0.005)***	(0.014)***	(0.068)
Share of Hispanic students	-0.003	-0.001	0.020
	(0.003)	(0.009)	(0.049)
Share of Female students	-0.001	-0.004	0.019
	(0.003)	(0.008)	(0.043)
Share of Students with Limited English Proficiency	-0.004	0.008	0.019
	(0.002)*	(0.007)	(0.030)
Share of Economically Disadvantaged Students	-0.068	-0.058	-0.081
	(0.004)***	(0.013)***	(0.056)
Share of Students with Disabilities	-0.011	-0.026	0.051
	(0.004)***	(0.010)**	(0.051)
Share of Academically Gifted students	0.017	0.035	0.010
	(0.003)***	(0.010)***	(0.031)
Panel B- Student Test Scores			
Student Test Score: Math	0.105	0.167	0.116
	(0.016)***	(0.040)***	(0.177)
Student Test Score: ELA	0.103	0.107	0.025
	(0.022)***	(0.060)*	(0.279)
Number of Movers	6162	686	34

Notes: This table compares the characteristics of students taught by teachers who move ("Movers") before and after each move. Each column reports the difference between post-move and pre-move means for the same teacher, separately by move number. Panel A reports student demographic characteristics, and Panel B reports average student test scores, standardized within subject, grade, and year. Standard errors in parentheses are from two-sample t-tests allowing for unequal variances. While estimated changes are similar across moves, estimates for third moves are imprecise due to the small number of such transitions.

Table 6: Chetty, Friedman, and Rockoff (2014a) Test of Forecast Bias

	(1)	(2)	(3)	(4)	(5)
	Δ Student Avg. Residuals	Δ Student Avg. Residuals	Δ Student Avg. Residuals	Δ Student Avg. Residuals	Δ Student Avg. Residuals
Δ Teacher Value-Added	1.103***	1.008***	1.638***	0.794***	1.582***
	(0.0530)	(0.0635)	(0.127)	(0.135)	(0.154)
Subject	All	Math	Reading	Math	Reading
Grade Level	All	Elementary	Elementary	Middle	Middle
Observations	26195	7247	7249	5842	5857

Notes: This table implements the test of forecast bias developed in Chetty et al. (2014), and where a switcher is defined as in the original paper: to include moves that exploit both within-school and across school variation. Specifically, coefficients in this table come from regressions of the form $Y_{ist} = \alpha + \beta\hat{\mu}_{j(s,t)} + \gamma X_{it} + \delta_s + \lambda_t + \varepsilon_{ist}$, where Y_{ist} is student achievement, $\hat{\mu}_{j(s,t)}$ is are leave year out jacknife estimate of teacher value-added, X_{it} are student controls, and δ_s and λ_t are school and year fixed effects. Columns denote different subject-by-grade level samples. Standard errors are clustered at the level of the teacher.

on their pay code. The first two digits of the pay code contain information on the teacher's education and the next two indicate the teacher's number of years of experience. These are the two variables NCDPI uses to determine teacher pay. While the NCERDC provides its own measure of experience in the dataset, we find it to be less reliable than the pay-based measure. It is more likely to have discontinuous jumps in teachers' years of experience from one year to the next, and in some cases even negative changes in teachers' experience from one year to the next. Eight percent of teacher-years have irregular changes from year-to-year in the NCERDC years of experience as opposed to only three percent for the pay-code variable.[31]

When there are missing or abnormal values for experience in the pay-code, we can impute years of experience for the same teacher in other years when we *do* see the pay code. We impute experience from up to four lead or lag years, and only when we observe the teacher in all intervening years. The priority is to impute from the closest forward or lag year. For any teachers with discontinuous jumps in years of experience that remain after imputation, we drop teacher-years after the jump.

When there are multiple and conflicting pay-codes for the same teacher-year, we first try to validate with lagged or lead years for the teacher. If this doesn't work, we use the modal year of report. In cases where the conflicting values are only one year apart, we use the lower year. If conflicting reports remain for a teacher-year, we drop this entry. Figure shows the distribution of teacher years in the sample before and after imputation.

[31] These irregular changes could be coming from administrative delays in updating a teachers' experience credit or due to data errors.

Figure 11: Distribution of Years of Experience

