ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 24-935

Next-Generation Teacher Evaluation in Rural Missouri: Main and Moderated Effects on Student Achievement and Effects-to-Expenditure Ratios

Seth B. Hunter
George Mason University

Katherine M. Bowser
George Mason University

We extend teacher evaluation research by estimating a reformed evaluation system's plausibly causal average effects on rural student achievement, identifying the settings where evaluation works, and incorporating evaluation expenditures. That the literature omits these contributions is concerning as research implies it hinders evidence-based teacher evaluation policymaking for rural districts, which outnumber urban districts. We apply a difference-in-differences framework to Missouri administrative data. Missouri districts could design and maintain reformed systems or outsource these tasks for a small fee to organizations like the Network for Educator Effectiveness (NEE), an evaluation system created for rural users. NEE does not affect student achievement on average but it improves math, and possibly reading, achievement in rural schools where the average student's prior-year achievement score is below the state average or the average teacher's years of experience are below the state average.

VERSION: March 2024

Suggested citation: Hunter, Seth B, and Katherine M. Bowser. (2024). Next-Generation Teacher Evaluation in Rural Missouri: Main and Moderated Effects on Student Achievement and Effects-to-Expenditure Ratios. (EdWorkingPaper: 24-935). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/x36v-vs97

Next-Generation Teacher Evaluation in Rural Missouri: Main and Moderated Effects on Student Achievement and Effects-to-Expenditure Ratios

Seth B Hunter, Corresponding Author
ORCiD 0000-0002-3051-872X
George Mason University
Assistant Professor of Education Leadership
4400 University Dr, MS4C2
Fairfax, VA 22030

Katherine M Bowser
ORCiD 0009-0004-4771-2272
George Mason University
PhD Candidate
4400 University Dr
Fairfax, VA 22030

Acknowledgements: The authors would like to thank the Network for Educator Effectiveness leaders, the Missouri Department of Elementary and Secondary Education, and participants from George Mason's education policy workshops and the Association for Education Finance and Policy for their helpful feedback.

Introduction

As of the mid-2020s, nearly every state education agency has implemented teacher evaluation reforms that include revised standards-based rubrics to assess teacher performance, changes to tenure, and frequent, structured performance feedback conferences (Bleiberg et al., 2024). According to education agencies, these reforms aim to improve teacher effectiveness via development primarily and accountability secondarily (Almy, 2011; Donaldson, 2021). As students taught by more effective teachers experience better short- and long-term academic and non-academic outcomes, strengthening teacher performance is laudable (Chetty et al., 2014; Jackson, 2018; Kraft, 2019; Liu & Loeb, 2021). Furthermore, improving the performance and productivity of the least effective teachers is a matter of equity as students from marginalized groups are systematically assigned to these teachers (Clotfelter et al., 2006; Kalogrides & Loeb, 2013). However, the start-up and ongoing costs of popular teacher evaluation reforms can be expensive (Chambers et al., 2013; Stecher et al., 2016) and may impose substantial burdens on school administrators (Kraft & Gilmour, 2016a; Rigby, 2015). These potential costs and benefits underscore the importance of examining evaluations' effects on student outcomes.

Despite the widespread adoption and importance of evaluation reforms, rigorous quantitative research examining evaluations' effects on student achievement is thin.[1] We have learned a lot about teacher evaluation in a few urban centers (Dee & Wyckoff, 2015; Steinberg & Sartain, 2015; Taylor & Tyler, 2012) and one national study (Bleiberg et al., 2024). Findings suggest that evaluations' effects on student achievement range from near-zero to substantially positive, though most effects are near-zero. Importantly, two studies also examine school

[1] However, a larger body of work examines evaluations' effects on other outcomes including teacher mobility (Cullen et al., 2021; Rodriguez et al., 2020) and student office referrals (Liebowitz et al., 2022). A multi-site randomized control trial also identifies the effect of providing educators with performance feedback, one aspect of RTTT-inspired evaluation reforms, on student achievement scores (Song et al., 2021).

characteristics moderating these effects, enabling targeted policy implications; one finds that the magnitude of positive effects increases with teacher years of experience (Taylor & Tyler, 2012), while another concludes that evaluations' positive effects increase with school-level student economic advantage and prior-year achievement scores (Steinberg & Sartain, 2015); both suggest that evaluation reforms generate Matthew effects.

However, no quasi-experimental research focuses on the costs or effects of teacher evaluation on student achievement in rural settings, a grossly underexamined context in teacher evaluation research, although most districts and school boards are rural (The School Superintendents Association, 2017). Emerging research finds education policymakers prioritize generalizability over internal validity (Nakajima, 2022); thus, the absence of rigorous, rurally situated teacher evaluation studies has left those who craft policy affecting rural schools in the dark if findings from nonrural settings do not generalize. Prior conclusions warranting causal inference may not generalize to rural settings for two reasons. First, the initial and ongoing research, development, and support for the implementation of reformed teacher evaluation systems, including teacher performance rubric and measure design, professional development for principals (teacher evaluation's primary implementers), and maintenance of a performance data management system, requires education agency time, capacity, and resources that may be prohibitive for rural districts (Chambers et al., 2013; Stecher et al., 2018; Gilles, 2017). Second, weak rural teacher labor markets may dissuade principals from dismissing underperforming teachers, suppressing a core mechanism by which teacher evaluation aims to improve the distribution of teacher effectiveness (Rodriguez et al., 2020). Furthermore, if principals avoid providing teachers with critical performance feedback following classroom observations to ensure they remain in their school, principals are foregoing practically large feedback-induced

teaching improvements (Hunter & Steinberg, 2022). These conditions suggest that urbanicity may moderate evaluations' effects on student outcomes. Additionally, no rurally focused study links these effects to expenditures, information policymakers need to make informed decisions.

Ultimately, there is insufficient evidence for the scientific community to reach defensible conclusions about the costs and effects of recent teacher evaluation reforms on rural student achievement scores and even less evidence regarding the conditions in which these reforms improve rural student outcomes; this paper addresses these gaps by answering the following research questions:

1. How much do districts spend to outsource the development and maintenance of a developmentally focused teacher evaluation system?
2. What is the impact of introducing an evaluation system on student math and reading scores in rural settings?
3. To what extent do these effects vary by school level: a) average years of teaching experience, b) FRPL concentration, c) nonwhite student concentration, and d) average student prior year experience achievement scores?

We investigate a unique teacher evaluation system, the Network for Educator Effectiveness (NEE). While prior work focuses on teacher evaluation systems managed by state departments of education or district central offices,[2] an independent center at the University of Missouri manages NEE. Notably, districts join NEE voluntarily and pay an annual fee to cover ongoing operational costs. Despite the availability of a state-designed teacher evaluation system without fees and the option for districts to design their own system, the number of districts

[2] Researchers designed the feedback intervention studied by Song and colleagues (2021).

choosing NEE has steadily grown from six of Missouri's 500+ districts in 2011-12 to 320 districts in 3 states presently, underscoring the relevance of examining this system.

NEE purposefully addresses the two conditions that we argue may moderate teacher evaluation's effects by urbanicity. It helps districts cope with the human, financial, and physical resource constraints in rural settings by developing and maintaining a theoretically sound teacher evaluation system for a remarkably low cost due to economies of scale and its location inside a public university system, which offsets some of NEE's operational costs. NEE is also sensitive to the reality of rural teacher labor markets and therefore eschews teacher evaluation's accountability mechanism completely. Consequently, NEE focuses on teacher development exclusively and effectively expands the developmental capacities of rural central offices.

To answer our research questions, we compare NEE's fee per student to other documented teacher evaluation costs and apply quasi-experimental methods to five years of panel data to estimate NEE's average effects on rural student achievement. To complement the existing literature that suggests teacher evaluation has not improved student achievement on average, we also estimate heterogeneous effects in policy-relevant school settings. Our average quasi-experimental estimates are precisely estimated, negligibly positive, and statistically insignificant, resembling findings from other settings. However, the data repeatedly suggest that NEE student achievement scores increase in schools where the average student's prior-year achievement score and the average teacher's years of experience are below state averages, contrary to the Matthew effects in non-rural settings. District expenditures suggest that NEE's annual costs to districts are among the cheapest of any documented system; consequently, effects-to-expenditure ratios in settings with positive estimates are remarkably high. Our findings

imply that rurally concerned policymakers might incentivize disadvantaged schools to adopt a teacher evaluation system resembling NEE.

Background

Theory of Action Framing Reformed Teacher Evaluation Systems

Theoretically, reformed teacher evaluation systems improve teacher effectiveness through two mechanisms: teacher accountability*,* which results in the forced or voluntary exit of ineffective teachers from the teacher workforce, and teacher professional development (PD), which improves individual teacher effectiveness (Donaldson, 2021; Papay, 2012; Phipps & Wiseman, 2021). The accountability mechanism operates through several sub-mechanisms. Reformed systems include standards-based teacher performance criteria and rubrics, and the higher frequency of classroom observations and post-observation feedback conferences allow evaluators to clarify these expectations (Donaldson, 2021). Conceptually, teachers who persistently struggle to meet expectations will be dismissed or exit the teacher workforce voluntarily, increasing student achievement as students gain access to more effective teachers (Donaldson, 2021); however, evidence supporting this hypothesis is mixed (Cullen et al., 2021; Rodriguez et al., 2020). Alternatively, performance accountability may motivate teachers to improve their teaching, ultimately improving student achievement (Phipps & Wiseman, 2021).

The developmental components of evaluation reforms might also improve teaching quality independent of pure accountability mechanisms. Observation conferences can provide teachers with performance-enhancing strategies. As reformed systems include higher frequencies of observations and post-observation feedback conferences (Steinberg & Donaldson, 2016), teachers effectively receive higher dosages of performance feedback. Notably, the feedback itself may not improve teaching directly (Cherasaro et al., 2016; Ilgen et al., 1979; Murphy &

Cleveland, 1995). Instead, feedback may lead teachers to PD opportunities tailored to observation-identified areas of weakness (e.g., targeted coaching; Donaldson, 2021), underscoring the importance of linkages between evaluation and PD systems (Kraft & Gilmour, 2016b; Weisberg et al., 2009). Ultimately, evaluation as a developmental tool theoretically depends on feedback quality, pointing towards the significance of evaluators' observation and feedback skills (Hunter & Springer, 2022; Kimball & Milanowski, 2009).

Teacher Evaluation in Rural School Settings

Due to a smaller labor pool, rural teacher labor markets are less elastic than non-rural markets, and rural schools face greater challenges in recruiting and retaining teachers than their urban peers (Nguyen et al., 2020). Furthermore, sparsely populated states (including Missouri, the setting of the study herein) experience higher rates of novice teacher turnover in rural communities than in urban centers, effectively placing rural schools in a constant state of onboarding and development (Nguyen, 2020). Therefore, evaluation's accountability mechanism may not function as intended in rural settings. Indeed, Rodriguez et al. (2020) found that introducing a reformed evaluation system in Tennessee led to increased retention of more effective teachers and increased turnover of less effective teachers in urban but not rural settings.

Rural teacher labor market inelasticity may also affect teacher evaluation's developmental function. Prior work finds that teacher performance improves when observers provide more critical feedback; however, receiving more critical feedback may (unintentionally) result in teacher turnover as teachers leave critical feedback settings and move into positive feedback schools (Feng, 2010; Hunter & Springer, 2022; Hunter & Steinberg, 2022). Thus, rural school leaders may forego critical feedback and more severe observation ratings that could improve teacher performance to retain teachers. Indeed, Hunter and Steinberg (2022) find that

principals consistently provided positive feedback following historically low-performing teachers' final observation of the year; the authors theorize that principals do this to increase the odds that these teachers will not leave their school.

District human, financial, and physical resource constraints may also affect teacher evaluation in rural settings. The financial barriers to implementing reformed evaluation systems are significant as rural schools may not be able to afford the infrastructure (e.g., data management systems) nor the personnel needs (e.g., training) required to implement a reformed evaluation system effectively. For example, without the data management system or central office staff support many urban districts possess, an already time-consuming teacher evaluation process would become even more burdensome and unmanageable for rural school leaders, impairing implementation.

Related Studies Regarding Teacher Evaluation Effects

We focus on the causal effects of introducing reformed teacher evaluation systems on student achievement scores, which only a few studies examine.[3] In a unique randomized control trial, Steinberg and Sartain (2015) estimated the effects of a reformed teacher evaluation pilot, the Excellence in Teaching Project (EITP). EITP, a low-stakes system that did not relax teacher tenure protections, was implemented across two cohorts of elementary schools in Chicago Public Schools. While analyses of student math scores did not detect effects, student reading scores increased significantly, although these effects were concentrated in the first EITP cohort. Notably, Cohort 2 schools did not receive the same administrative and implementation support as

[3] A larger body of work estimates the effects of related but dissimilar treatments on student achievement scores or teacher value-added to achievement scores. For example, Dee and Wyckoff (2015) identify the effects of evaluation-triggered (dis)incentives, and Song and colleagues (2021) estimate the effects of providing educators with performance feedback measures. As these treatments differ from the treatment of introducing a next-generation evaluation system, we do not discuss them further.

Cohort 1 schools. This is the only study we know of that estimates the heterogeneous effects of school-level characteristics; in broad terms, advantaged schools (i.e., higher-performing and lower-poverty) benefited more than disadvantaged schools. There was no evidence of heterogeneity by school-level shares of student race or average teacher years of experience.

A quasi-experiment by Taylor and Tyler (2012) examines the impact of a next-generation evaluation system implemented in Cincinnati Public Schools. Specifically, the authors analyzed the impact of next-generation evaluation on mid-career teachers' students' achievement scores. While reading scores were unaffected, student math scores increased significantly in the years after a teacher went through the evaluation cycle. These results were concentrated among teachers in the bottom half of the distribution of prior evaluation scores.

A recent nationwide study using data from the Stanford Education Data Archive found that reforms had, on average, a null effect on student achievement (Bleiberg et al., 2024). The authors also examined heterogeneity across system design features by constructing an index of design rigor (e.g., high-stakes, bonus pay, observations required) and found precise null effects across indices. Notably, the authors hypothesized that ineffective implementation explains the lack of student achievement gains.

Related Cost Studies

Two studies examine the per pupil expenditures associated with teacher evaluation systems and both use data from the Intensive Partnerships for Effective Teaching (IP) initiative funded by the Bill & Melinda Gates Foundation. Seven IP districts – three school districts and four charter management organizations – implemented reformed teacher evaluation systems from 2009-10 through 2014-15. The systems examined included reforms shared by NEE, EITP, and the Cincinnati systems. The overall per-pupil expenditures ranged from $868 to $3,541 (Stecher

et al., 2018). For five of the seven sites, one-time bonuses and permanent salary increases tied to teacher evaluation were the largest source of direct expenditures, while two sites spent the largest portion of their funding on principal and teacher PD. Additionally, Stecher et al. (2018) determined that accounting for the time teachers and school leaders spent on evaluation would add $200 per pupil, on average.

Chambers, Brodziak, and O'Neil (2013) ("CBO") provide a disaggregated cost analysis for the first three years of implementation at the three traditional school district IP sites. First, CBO separate per-pupil expenditures by the three broad system components – costs regarding teacher observation, student surveys, and value-added measures. Costs are further disaggregated by start-up and ongoing costs and five subdomains. The 'design and implementation' subdomain included costs associated with developing materials and procedures (e.g., designing an observation rubric, training observers). A 'peer, mentor, or external evaluators' subdomain included the salaries and benefits of those hired solely to conduct observations. The 'management and communication' subdomain captured expenditures related to the resources needed to introduce reforms to district staff, including teachers and principals. A 'technology and data systems' subdomain captured expenditures related to developing or purchasing a central performance data system, and the 'other' subdomain included all unassigned costs. The authors defined start-up costs as one-time expenses related to designing and planning the systems, while ongoing costs were regularly occurring (e.g., annual) tied to operating and maintaining the elements of the teacher evaluation system.

CBO found that the per pupil yearly expenditures ranged from $8 to $118 across system components over the first three years of implementation. Each district spent the highest proportion (47%-87%) of funds on the teacher observation component, though the subdomains

driving these expenditures varied across districts. One district spent a great deal on 'peer, mentor, and external evaluators' while the remaining districts relied almost exclusively on principals and assistant principals to conduct observations – as do NEE districts. The second most expensive subdomain for each district was 'management and communications' followed by 'technology and data.' Importantly, CBO likely underestimated the total costs for two reasons. First, the expenditures do not include the time spent by district personnel (who were not hired explicitly to manage the reformed system). Second, the districts examined began prepping for the reformed evaluation systems before the study period; thus, some costs may have occurred before the study period.

Study Context

In the early 2010s, researchers at the University of Missouri's College of Education developed NEE with substantial input from recently retired PreK-12 rural principals and district administrators. NEE recruited six districts into Cohort 1 to pilot the system in 2011-12 and launched training during the summer of 2011. Notably, Cohort 1 districts did not know about the summer 2011 launch until spring 2011, and districts did not receive their 2010-11 student achievement scores until after recruitment. These conditions mitigate concerns about the endogenous timing of NEE's adoption and anticipatory effects, particularly those arising from student achievement scores and their correlates. NEE recruited 26 districts into Cohort 2 for a 2012-13 launch and trained those districts in the summer of 2012 before participants received 2011-12 achievement scores. Recruitment in both cohorts was based entirely on district urbanicity and NEE developers' professional networks with rural superintendents. All recruited districts joined, and superintendents treated NEE as a district-level policy that they expected all their schools to implement.

NEE Design Element 1: Observation Rubrics and Goal Setting

NEE includes an observation rubric describing research-based instructional practices aligned with Missouri's teacher performance standards and resembles Danielson's ubiquitous Framework for Teaching (Marshall, 2013). NEE's rubric includes 39 teaching indicators measuring nine Standards of Teaching. Each indicator defines five performance levels (0, 1, 3, 5, 7), though teachers can receive any integer score 0-7. Several studies validate the NEE rubric (e.g., Bergin et al., 2017; Wind et al., 2018).

Setting individual performance goals using performance rubrics is a theoretically essential component of any evaluation system (Choi & Johnson, 2022; Locke & Latham, 2002); indeed, recent work suggests rubric-based goal setting may be one of the main mechanisms by which evaluation improves early-career teacher performance (Hunter & Springer, 2022). NEE teachers actively engage in goal-setting processes with school administrators to select annual teacher performance goals, suggesting that this design element can improve student achievement.

NEE Design Element 2: Observation Frequency and Conferences

Theoretically, classroom observations are the linchpin of evaluation for development as they can include teaching performance assessment, goal-setting, performance-enhancing feedback, and improvement plans (Donaldson, 2021). Empirically, the effects of more observations on student achievement vary by context. Research from DC's high-stakes evaluation system suggests that the marginal observation improves teaching and student achievement (Phipps, 2022; Phipps & Wiseman, 2021). However, larger-scale research from more typical low-stakes settings finds no effects on student achievement (Hunter & Kho, Online), although policy-assigned observations reduce student exclusionary discipline outcomes (e.g., out-of-school suspensions; Hunter et al., 2023). Combined with the aforementioned

findings from Hunter and Springer (2022), we interpret the evidence and theory to mean that growth is more likely when teachers are observed more frequently.

NEE recommends that every teacher receive six to ten mini-observations per year. During the study period, we do not know how many observations were received per teacher; however, prior work from other settings finds that principals typically conduct fewer observations than teachers are assigned (Hunter & Ege, 2021; Hunter & Kho, Online; Kraft & Gilmour, 2016a). Indeed, NEE observation data collected after the period examined by the study herein reveals that the typical teacher received four yearly observations (Bowser & Hunter, 2022). Though this is below the NEE recommendation, it is an unusually high number of observations relative to other systems and may result in greater gains in student achievement (Hunter, 2020; National Council on Teacher Quality, 2019).

NEE Design Element 3: Observer Preparation and Certification

NEE evaluators receive annual and ongoing training and support to promote reliable and accurate scoring. Evaluators also receive training about how to provide performance feedback effectively. Training also focuses on collaboration with teachers directly and supporting teacher collaboration with other personnel (e.g., peer mentoring) to improve observation-identified areas for improvement. Following training, prospective evaluators must pass a certification exam each summer to receive certification to conduct formal observations. Theoretically, design element three should also increase the odds that NEE improves student achievement.

NEE Fees

NEE charges districts an average of $3 per student to cover its operational costs. We compare NEE's fee to the per pupil expenditures detailed in CBO.[4] In NEE's first two years (the period examined herein), the $3 per pupil fee included access to NEE's observation scoring rubric, the NEE Data Tool (a centrally managed performance data system), evaluator (school administrator) initial certification and yearly recertification training sessions, ongoing training for educators, webinars, and technical support via the NEE Help Desk.

Data

This study uses grades 3-8 statewide administrative data from Missouri's Department of Elementary and Secondary Education (DESE), NEE-supplied lists of its first two cohorts, and National Center for Education Statistics (NCES) urbanicity and per-pupil expenditures (PPE) from 2007-08 through 2012-13. DESE allows the linkage of schools to districts, students to schools, and teachers to schools. Student administrative data includes race, gender, FRPL, and achievement scores, while teacher data includes race, gender, education level, and years of experience. As NEE is fee-based and designed for rural districts, we control for urbanicity and PPE via NCES data.

NEE adoption is a district-level policy; thus, our independent variable is at the district level. Our outcome variable is student-level math and reading achievement. While the number of NEE districts is small, our sample size is sufficiently large. Cohort 1 included 24 schools enrolling approximately 5,000 students, and Cohort 2 included 71 schools enrolling approximately 10,000 students.[5]

[4] We use costs reported in CBO due to the detailed disaggregation of expense categories. Stecher et al. (2018) does not disaggregate costs to the same extent as CBO, making comparison to NEE less clear. Further, CBO reports costs of districts in the early years of policy adoption, similar to the study herein.

[5] For context, the data in Steinberg and Sartain (2015) included 44 schools in its first cohort and 48 in the second, and Taylor and Tyler (2012) included a rough total of 3,600 treated and untreated students.

Methods

Cost Analysis

To answer our first research question, we use the disaggregated cost analysis of CBO to construct relevant comparisons to NEE's fee. First, we align the components defined in CBO with NEE's products and services. Then, we identify conservative to liberal ranges of ongoing costs[6] that a district would have had to spend to implement a reformed teacher evaluation system.

Quasi-Experimental Analysis

NEE was not assigned to districts randomly; however, we use our strong knowledge of the selection process and NEE's discontinuous rollout to estimate the causal impact of NEE's introduction on changes in student achievement. Although some recent research eschews generalized DDs with two-way fixed effects due to concerns about heterogeneity arising from differential lengths of treatment exposure (Callaway & Sant'Anna, 2021; Goodman-Bacon, 2021), these concerns are unwarranted in our case for two reasons. First, we estimate cohort-specific effects one year after treatment to examine cohort heterogeneity and one-year-after effects pooled across cohorts. Second, when we estimate effects for NEE's first cohort only, which was exposed to NEE for two years, we estimate cohort-year-specific effects.

We apply generalized DD designs, which assume that the changes in achievement scores among the treated would have been statistically equivalent to achievement score changes for the comparison group had the treated not been exposed to NEE. This supposition, known as the parallel trends assumption, is violated if time-varying omitted variables (OVs) affect the treated and comparison groups differentially and substantively in the pre- or post-treatment periods. While observed achievement scores allow us to estimate if achievement scores changed in

[6] For completeness, we complete a similar comparison for start-up (i.e., one-time) costs.

treated versus comparison groups differentially and substantively in the pre-treatment period, we do not observe post-treatment counterfactuals. However, limiting our analyses to two years after treatment, at most, decreases the chance that unobserved concurrent or post-exposure alternative treatments (i.e., OVs) affected student achievement trends in NEE districts but not comparison groups. Moreover, this short time frame, combined with the relatively small number of districts in Cohorts 1 and 2, means NEE leaders were aware of alternative treatments in their districts. Over two interviews, NEE leaders emphatically reported that NEE districts did not receive alternative treatments, bolstering the plausibility of parallel trends post-treatment.

Selection

NEE leaders also repeatedly stated that districts were recruited on two conditions: if the district was rural and its superintendent had a strong professional relationship with NEE leaders. NEE leaders also noted stronger ties with the superintendents in NEE Cohort 1 than Cohort 2.

We observe urbanicity but not NEE's professional ties with superintendents; however, we argue that it is implausible for the professional ties between superintendents and NEE leaders to represent an OV capable of undoing our causal inferences. We accept that NEE leaders' relationships with superintendents might have been influenced by their perceptions of superintendent effectiveness, but we argue that selection on perceived superintendent effectiveness is not concerning. NEE would have needed to select on achievement-related superintendent effectiveness (instead of, for example, political acumen) to raise the possibility that treated achievement scores changed because of superintendent effectiveness. Not only did NEE leaders emphatically deny that student achievement affected their relationships with superintendents, but research concerning superintendent effects on student achievement suggests

that such effects are relatively small and, therefore, unlikely to threaten our inferences (Schwartz et al., 2023).

Nonetheless, we empirically assess the potential threat of superintendent relationships or effectiveness in two ways. NEE leaders' relationships with superintendents in Cohort 1 were stronger than their relationships with superintendents in Cohort 2, which were stronger than relationships with non-NEE districts. Given the monotonicity of relationship strength across Cohort 1, 2, and non-NEE districts, Cohort 1 versus non-NEE comparisons should produce larger estimates than Cohort 1 versus Cohort 2 comparisons *if superintendent relationships explain our estimates*. Second, we examine the conditions in which any OV could undo our inferences; the evidence repeatedly suggests that such conditions are implausible.

Matched and Stacked Generalized Difference in Differences

We define treatment as exposure to NEE implementation, and the comparison group consists of students and schools in districts that did not implement NEE. We apply a matching procedure to identify the non-NEE districts resembling NEE regarding PPE in the year before treatment and district-level average student achievement scores one, two, three, and four years before treatment. We match on historical achievement trends and prior-year PPE as some may believe these variables must have affected selection into a fee-based teacher evaluation system aiming to improve student achievement, contrary to the emphatic accounts of those who worked with NEE districts closely over several years.

Nonetheless, we identify matched comparison districts using coarsened exact matching (CEM) per Sturge's Rule, in which districts are the unit of analysis since selection was at that level. Matching occurs by cohort; the pool of potential matches for Cohort 1 includes all rural districts that did not implement NEE through 2011-12, the year Cohort 1 launched. Districts that

implemented NEE in 2012-13 were also in Cohort 1's pool of potential matches for 2011-12; Cohort 2 is omitted from the pool of potential matches for 2012-13. Cohort 2's matching procedure is analogous to Cohort 1's, except that the pool of potential matches includes all rural districts that did not use NEE through 2012-13. CEM matches on five variables: district-level PPE from the year before treatment and district-level average student achievement scores one, two, three, and four years before treatment.

Matching on district-level average student achievement scores one year before treatment, and the timing of NEE's recruitment bolsters the parallel trends assumption considerably. Student achievement scores from the spring before NEE's launch capture variation in all (unobserved) factors that determined those scores up to the point of assessment. If NEE selection occurred after prior-year spring assessments, selection could have been affected by OVs that student achievement scores did not capture; however, NEE recruitment (selection) occurred before achievement testing in the spring before launch. These conditions suggest that matching on prior-year scores alone effectively controls for the OVs affecting selection and prior-year achievement. While our model does not control for factors affecting selection independent of prior-year achievement scores, such factors could only threaten the parallel trends assumption if they determined selection and post-treatment achievement scores but not pre-treatment scores from any of the four prior years; we assume such factors are implausible.

After identifying district matches for Cohorts 1 and 2, matched data are returned to the student level and stacked; Cohort 1 and its matches are stacked onto the data for Cohort 2 and its matches, yielding a student-year-cohort dataset. Years within each cohort/ stack are centered on NEE's introduction year (e.g., Cohort/ Stack 1 year 0 corresponds with 2011-12) and range from

-4 to 0. Following Gormley and Matsa (2011), we apply a generalized DD model to stacked data using Equation 1:

$$y_{isdtc} = \delta NEE_{dt} + \beta_1 y_{isd(t-1)} + \beta_2 PPE_{d(t-1)} + \tau_{dc} + \pi_{tc} + e_{isdtc} \quad (1).$$

Where y_{isdtc} is the grade-standardized math or reading achievement score of student *i* in school *s* in district *d* in centered-year *t* in cohort *c*. NEE_{dt} is the treatment variable. Equation 1 applies district-cohort fixed effects (FE) and year-cohort FE, effectively comparing achievement trends within each stack (Gormley & Matsa, 2011). Equation 1 also includes prior-year student achievement and district PPE; we do not control for urbanicity because we limited the sample to rural districts only. We focus on changes in student achievement scores one year after the pilot as this limits the probability of post-treatment threats to the parallel trends assumption and follows the one-year-after treatment estimates in most related work (i.e., Steinberg & Sartain, 2015; Taylor & Tyler, 2012). We also estimate changes in student achievement two years after Cohort 1's pilot. Our preferred specification uses standard errors that are district-student-cohort multiway clustered.

Sensitivity Tests

Our sensitivity tests begin by re-applying Equation 1 with an augmented set of control variables. If adding augmented controls to Equation 1 results in substantively different inferences, it could indicate a violation of the parallel trends assumption. While we can control for observable differences, the sensitivity of estimates generated by Equation 1 to the augmented model would raise concerns that effects attributed to NEE result from unobserved OVs. The augmented controls include student race, gender, FRPL, and the proportion of students in a school and district by race, gender, and FRPL; the concentration of teachers in a school and district by race, gender, education level, and average teacher years of experience; and school-

and district-level average student prior-year achievement scores. We also estimate versions of Equation 1 using i) the canonical district FE and year FE, ii) district FE, year FE, and augmented controls, and iii) district-cohort FE, year-cohort FE, and cohort-specific augmented controls.

We also use Rosenbaum and Rubin (1983) sensitivity tests to estimate the degree of potential bias resulting from an omitted variable (OV); this test has been further developed by Cinelli and Hazlett (2020) to report the maximum bias of multiple, non-linear confounders. Notably, correlations between OVs and residual outcome and treatment variation (i.e., variation not explained by the model) would not be sufficient to undo inferences if the OV it would take to do so is implausible; analysts must explain why the reported confounding conditions are not plausible. We determine what is plausible using the explanatory power of the observed covariates that strongly determine outcome variation. We compare the hypothetical OV against prior-year student-level and achievement scores and argue that it is implausible that an OV could explain more achievement-score residual variation than what is explained by this variable.

Heterogeneity

We explore heterogeneity by school characteristics and cohort; the latter also serves as a sensitivity test assessing if unobserved between-cohort differences affect our inferences (e.g., the strength of superintendent relationships). We create the means of the following school-characteristic moderators: average teacher's years of experience (=12), proportion of nonwhite students enrolled in a school (=10%), proportion of FRPL students enrolled in a school (=50%), and school-level average student prior-year achievement score (=0). Next, we create four binary moderators, each taking a value of zero if the school is below the state average and one if it is at or above and interact the moderators with NEE_{dt}.

Parallel Trends Test and Event Study Analysis

Event study analysis explores pre-intervention parallel trends and estimates treatment effects nonparametrically. The event study analysis compares pre- and post-intervention student achievement in NEE and matched non-NEE districts by each year preceding NEE's launch and the year of its launch in each cohort. Equation 2 describes the event study model:

$$y_{isdtc} = \delta_{-4}NEE_{dt} + \delta_{-3}NEE_{dt} + \delta_{-2}NEE_{dt} + \delta_0 NEE_{dt} + \beta_1 y_{isd(t-1)} + \beta_2 PPE_{d(t-1)} + \tau_{dc} + \pi_{tc} + e_{isdtc} \quad (2)$$

Equation 2 replaces δNEE_{dt} with interactions of year dummies and treatment status, omitting the interaction between the year preceding NEE's launch and treatment status; consequently, δ_j represents the difference in achievement scores j years before or after NEE's launch relative to the difference in the year preceding NEE. If achievement trends in NEE and matched non-NEE districts are relatively parallel over time, meeting a DD identification assumption, δ_j will be statistically insignificant when $j < 0$. Additionally, δ_0 corresponds with Equation 1's δ; other terms refer to the same quantities as Equation 1.

Placebo Tests and Effects on Covariates

Estimates may be biased if interventions in the years preceding NEE's launch affected student achievement later. Placebo tests estimate these pre-NEE 'effects' using false NEE launch dates. Specifically, the first placebo test recodes Equation 1's NEE_{dt} so it equals one for NEE districts in the year preceding NEE's launch and thereafter (e.g., Cohort 1 year $\geq$ 2010-11; centered-year $\geq$ -1). The remaining placebo tests similarly recode NEE_{dt} for the remaining false years of treatment.

We have no reason to believe that NEE's introduction should affect the observable compositions of NEE districts or any other covariate. Indeed, if treatment 'affects' covariates, it

could suggest alternative treatments. Online Appendix A describes the baseline balance tests in detail.

Effects Over First Two Years: Cohort 1

Although the study's primary purpose is to identify the effects after one year of implementation, Cohort 1's data permit estimating NEE's effects one and two years after introduction. We only retain Cohort 1 and its matched comparison group to estimate these dynamic effects. Cohort 1 and its matched comparison group data from 2012-13, its second year of implementation, are also added to the sample. As the new sample is not stacked, district-cohort FE and year-cohort FE are replaced with district FE and year FE. We estimate dynamic post-intervention effects by adding an interaction to Equation 1, interacting NEE_{dt} with an indicator marking if the records came from 2012-13.

Findings

Cost Comparison

Table 1 displays cost ranges from CBO. The lower bound includes only those costs from CBO subdomains with clear connections to NEE, while the upper bound includes all teacher observation-related costs. Recall that CBO accounted for teacher observation, student survey, and value-added measure (VAM) cost domains. We only discuss teacher observation-related costs since NEE did not utilize VAMs or surveys during the study period.

As discussed above, CBO identified five teacher observation cost subdomains: i) design and implementation; ii) peer, mentor, and external evaluators; iii) management and communications; iv) technology and data systems; and v) other. We argue that CBO's (i), (iii), and (iv) subdomains clearly connect to NEE's services (e.g., observation rubric and related resources, observer training and calibration, a data management system, and technical support).

Although we argue that NEE did not include expenditures concerning (iv) and (v), we add these to the upper bound costs for reference. Definitions of (i) – (v) and rationale for NEE's aligned services are detailed in Online Appendix Table B1. Finally, we adjust CBO costs to 2012 real dollars as NEE was launched in 2011-12.

While we show CBO ongoing costs for each of the three years the authors examined, we assert that Year 3 costs are the best comparison; ongoing costs in Years 1 and 2 are low and at times near zero because districts are in a start-up phase. During this phase, districts are engaging in planning activities (i.e., one-time start-up expenses that CBO does not include in the ongoing cost reports). NEE districts do not undergo this start-up phase, however for completeness we also provide start-up costs in Appendix Table B2. Focusing on the lower bound Year 3 costs in Table 1, we see that the lowest CBO cost is approximately $6 per pupil, or twice NEE's fee. The upper bound estimates suggest that districts may spend as much as $50 per pupil, or more than 16 times NEE's fee.

Pre-Matched Descriptive Statistics

NEE districts resemble the sample of all non-NEE districts in several ways (Table 2). However, NEE districts enroll lower percentages of nonwhite students, all NEE districts are rural, whereas 16 percent of non-NEE districts are not, and the average NEE district spends about $1,500 less per pupil, countering the notion that districts choosing to pay NEE's nominal fee are wealthier. These differences underscore the need to adjust for differences between NEE and non-NEE districts.

Matching Results for DD Design

As the validity of our strategy does not depend on post-matching covariate baseline balance at the district level (it only depends on parallel trends and no alternative treatments), we

describe matching results briefly, beginning with the math score sample. Cohort 1 matching examined 234 coarsened strata and matched within four, matching five of six NEE districts to 67 non-NEE districts. Cohort 2 matching used 287 coarsened strata, matched using 16 strata and matched 19 of 26 NEE districts to 127 non-NEE districts. The mean differences between matched NEE and non-NEE districts across Cohort 1 and 2 districts ranged from -0.03 to 0.03 SD regarding prior-year average student math scores and -$250 to $195 in prior-year PPE.

Reading score matching resembles math sample results. Cohort 1 examined 168 coarsened strata and matched using four while Cohort 2 matching considered 207 coarsened strata, matching on 18. The matched reading sample differs from the matched math sample; five Cohort 1 districts matched 120 non-NEE districts, while 24 Cohort 2 districts matched 197 non-NEE districts. Mean differences between Cohort 1 and 2 matched reading groups ranged from -0.03 to 0.09 SD for prior-year average student reading scores and -$385 to $114 in prior-year PPE. Finally, each CEM procedure resulted in matched samples that only included rural districts (for further details, see Online Appendix C).

Descriptive District-Level Prior-Year Student Achievement Trends

There is some evidence that pre-intervention achievement trends in *pre-matched* districts that did not adopt NEE throughout the study period are not parallel to trends in districts that implemented NEE; however, graphical analysis suggests that the matching procedure successfully identified comparison districts with trends paralleling NEE district's prior-year student achievement scores. Figure 1 graphs the average district-level achievement scores in NEE, non-NEE, and matched non-NEE districts. The top-left panel suggests that pre-matched non-NEE and Cohort 1's pre-intervention math score trends are not parallel. While pre-matched non-NEE pre-intervention trends hover around -0.02, Cohort 1's ranges from approximately 0.08

to -0.05. However, the top-right panel shows that Cohort 2's pre-intervention math score trend parallels the pre-matched non-NEE trend. The matching procedure produced prior-year math score trends that parallel NEE trends in each cohort. Moreover, Cohort 1's trend and matched the non-NEE pre-intervention trend are near-equivalent. The bottom-left panel shows that NEE, all non-NEE, and matched non-NEE pre-intervention trends are largely parallel, although NEE district reading scores deviate from the trend four years before NEE implementation. Finally, the bottom-right panel suggests that Cohort 2 pre-intervention trends are parallel and near-equivalent.

Although Figure 1 suggests that district-level matching was successful, the parallel trends assumption of the DD design rests on parallelism in *student*-level pre-intervention trends, as students are the unit of analysis in the DD. We examine the parallelism of pre-intervention student-level achievement trends in NEE and matched non-NEE districts using event studies.

Post-Matching Generalized DD Results

NEE's average treatment on the treated (ATT) for math and reading scores are insensitive to model specification and not moderated by cohort. Table 3 shows that the ATTs on math and reading scores are 0.01 SD but not statistically significant (Column I). Equation 1's ATTs are not sensitive to the use of the expanded set of controls, cohort-specific controls, replacement of district-cohort FE and year-cohort FE with district FE and year FE, nor the use of the expanded controls with district FE and year FE (see Columns II – V). Indeed, the ATT is consistently 0.01 SD in each subject.

Given the near-zero and statistically insignificant ATTs on student achievement scores in both subjects, the rest of the paper focuses on the correlates of these effects. However, when

discussing the internal validity of subsample or moderated effects, we also discuss the internal validity of the sample that gave rise to Table 3.

Heterogeneity

Table 4 presents NEE's effects on student achievement, moderated by the binary moderators regarding school-level average student prior-year achievement and average teacher years of experience. The data suggest that NEE increased student-level math achievement by 0.04 SD in schools where the average student's prior-year achievement score was below the state average (Panel A, Column I). To understand if the data from these below-average schools warrant causal inferences, we test the internal validity of the research design for this subsample of schools below. We begin by re-estimating Equation 1 on the subsample of schools with below-average student achievement in math and find a similar effect (0.06 SD, Panel A, Column II). Notably, the subsample estimate in Column II is insensitive to the use of control variables (Panel A, Column III), suggesting that our research design already accounts for observable differences between NEE and non-NEE schools and unobservable differences that strongly correlate with the observables.

We repeat the moderation, subsample, and subsample with control variables analyses when examining other heterogeneous effects on math and reading scores. Table 4, Panel B suggests that NEE improved math scores in schools with below-average teacher years of experience (0.04 SD, Column IV), and this finding holds across the subsample and subsample with control variables analyses (Panel A Columns V – VI). Similarly, Panel B suggests that NEE improved reading achievement in schools with below-average prior-year student achievement (0.03 SD, Columns I – II); however, we lose precision after adding control variables to the model. Similarly, we find that NEE improves reading achievement in schools with below-

average teacher years of experience (Column IV), but this difference becomes statistically insignificant in the subsample analyses (Columns V – VI).

We do not detect any heterogeneous effects regarding the proportions of students in schools who are FRPL or nonwhite (see Online Appendix Table D1); consequently, we do not examine those heterogeneous effects further.

Parallel Trends Test and Event Study Results

Event study results show that pre-intervention achievement score trends are consistent with the parallel trends assumption, bolstering causal inferences, and suggest that NEE improved math achievement scores in schools with below-average prior-year student achievement and teacher years of experience and improved reading scores in schools with below-average prior-year student achievement (Table 5). Pre-intervention differences in achievement across NEE and non-NEE districts are statistically indistinguishable from the score difference in the year before NEE's launch – the omitted category. Panels A and B clearly show stable (parallel) differences in achievement scores between NEE and non-NEE districts in the below-average school subsamples until NEE's introduction, strongly suggesting that post-NEE differences are attributable to NEE. Furthermore, post-NEE event study estimates closely resemble the generalized DD estimates from Table 4. Specifically, among schools with below-average prior-year student achievement, NEE's introduction increased student-level math by 0.07 SD and reading by 0.03 SD relative to the year before treatment (Table 5, Panel A). Among schools with below-average teacher years of experience, NEE improved math scores by 0.05 SD relative to

the year before treatment, though its effect on reading achievement in these schools is near-zero and statistically insignificant (Table 5, Panel B).[7]

Between-Cohort Comparisons and Effects Over Time

Table 6 separates one-year-after-treatment effects by cohort, estimates one- and two-year-after-treatment effects for Cohort 1 only, and, for reasons described in the Selection section, compares achievement scores from Cohort 1 against Cohort 2 only.[8] Columns I and IV use the same samples as Table 5 but apply a version of Equation 1 where we interact a cohort identifier with the treatment indicator. Panel A Column I shows positive one-year-after changes in math achievement scores among schools with below-average student achievement across both cohorts; however, only Cohort 2's change is statistically significant. Among schools with below-average teacher years of experience, scores declined in NEE Cohort 1 and rose in Cohort 2, but only the latter is statistically significant (Panel B Column I). Effects on reading are somewhat similar. Among schools with below-average prior-year achievement scores, NEE's introduction increased reading achievement by 0.03 SD in both cohorts, but this time, each estimate is statistically significant (Panel A Column IV). Like the effects on math scores in schools with below-average teacher years of experience, there is a drop in reading achievement in Cohort 1 and a rise in Cohort 2; however, neither is statistically significant (Panel B Column IV). Notably, the evidence in Columns I and IV is inconsistent with positive selection on perceived superintendent effectiveness (see Selection section).

[7] We also apply event studies to each content-specific sample from Table 3, which includes below and above-average schools (see Online Appendix Table E1). Again, we detect no post-treatment ATTs, and the data corroborate the parallel trends assumption.

[8] See Online Appendix Table E2 for corresponding analyses using the full sample from which results in Table 3 were based.

Next, we turn to one- and two-year-after effects for Cohort 1; these analyses restrict the samples from Table 5 to only Cohort 1 and its matches and apply a version of Equation 1 in which we interact treatment with a variable indicating whether the data were collected one or two years after treatment (see Table 6 Columns II and V). Math scores in NEE Cohort 1 schools with below-average prior-year student achievement, compared to below-average non- or not-yet-NEE schools, suggest that NEE's effects on math scores may fade over time (Panel A Colum II). However, the opposite occurs among schools with below-average teacher years of experience (Panel B Column II): one year after NEE's introduction, math scores decline, though the change is statistically insignificant, and scores rise by a statistically significant 0.06 SD two years after NEE's introduction. While we know the precise type of human capital deficiency in schools with below-average teacher experience (i.e., on-the-job experience), schools with below-average prior-year student achievement may suffer from multiple human capital deficiencies. At face value, the results in Column II may suggest that NEE can immediately address the various levels of human capital needs in low-performing schools. At the same time, high concentrations of teacher inexperience may be less tractable, and therefore, schools with such teachers need more prolonged NEE exposure before realizing improvement. Results in Column V show that reading scores change in the same direction as math scores, though no estimates in either panel are statistically significant.

Finally, we examine results from another indirect test of positive selection by comparing NEE Cohort 1 schools to not-yet-treated NEE Cohort 2 schools only, leveraging variation in treatment timing (Columns III and VI). Recall that substantive positive selection on perceived superintendent effectiveness would cause achievement score differences between Cohort 1 and non-NEE schools to exceed those between Cohort 1 and 2 (see Selection section). Among

schools with below-average prior-year student achievement, the difference between Cohort 1 and non-NEE and not-yet-NEE math achievement is 0.07 SD (Table 4 Panel A Column III), while the difference between Cohort 1 and 2 is 0.09 SD (Table 6 Panel A Column III). The corresponding differences in math scores among schools with below-average teacher years of experience are 0.04 SD (Table 4 Panel A Column VI) and 0.07 SD (Table 6 Panel B Column III). These patterns are inconsistent with the assumptions regarding positive selection on perceived superintendent effectiveness, affirming the research design's internal validity concerning math scores. However, the evidence regarding reading scores is less compelling; the estimates in Table 4 Panel B Columns III and VI are small and statistically insignificant, and those in Table 6 Panels A and B Column VI are smaller and insignificant.

Sensitivity Analyses

Formal sensitivity tests suggest that our inferences are insensitive to plausible OVs, bolstering our confidence in the internal validity of our research design further. We examine OVs with up to 30% of the explanatory power of student-level prior-year achievement since an OV mimicking this relationship explains 95.4% of the residual variation in math scores in schools with below-average achievement (Table 7 Panel A Column II).

Table 7 Panel A presents sensitivity analyses regarding the effects of our hypothetical OVs on math and reading scores for those schools with below-average prior year student achievement. While none of the benchmarked OVs explain a substantial amount of residual treatment variation, student-level prior-year scores explain substantial amounts of residual outcome variation; however, an OV resembling this powerful predictor of residual outcome variation has virtually no effect on our coefficients or inferences (Table 7 Panel A Columns III and VI). We observe similar patterns among schools with below-average teacher years of

experience (Panel B Column III) and note that the effects of our OVs on reading scores, which have not resulted in significant effects in any model remain statistically insignificant.

Placebo Tests and Effects on Covariates

Placebo tests affirm the internal validity of the research design for estimating math score effects but present less compelling evidence for the design estimating reading score effects among schools with below-average prior-year achievement (Online Appendix Table E3). Panels A1 and A2 detect no placebo effects on math scores among either subsample of schools. However, results in Panel B1 asperse the internal validity of the research design for reading scores when applied to schools with below-average prior-year student achievement, though we detect no placebo effects on reading scores among schools with below-average teacher years of experience (Panel B2). Given the lack of detected effects on reading scores and evidence aspersing the internal validity of our research design when applied to reading subsamples, the remainder of the paper focuses on math scores. Online Appendix Table E4 displays placebo test results using the full samples from Table 3 and detects no placebo effects.

As a developmentally focused teacher evaluation system eschewing evaluation for accountability, NEE is not designed to alter the compositions of students or teachers in its schools or districts. Consequently, we should not detect any effects on variables describing these compositions; if we find compositional changes, it could suggest the presence of alternative treatments (that aim to change compositions). Online Appendix Tables F1 and F2 present no evidence suggesting that NEE introduced compositional changes among the math subsamples comprised of schools with below-average prior-year student achievement or teacher years of experience, once again authenticating the internal validity of our research design for math scores. However, we detect some compositional changes among schools with below-average prior-year

student achievement in reading (Online Appendix Table F1). While the number of changes we detect across Online Appendix Tables F1 and F2 could arise from Type I error, we conclude that the evidence collectively undermines the internal validity of our research design for detecting NEE's effects on reading scores. We also note that we detect no compositional changes among the full math and reading samples used in Table 3 (see Online Appendix Table F3).

Conclusion

There is insufficient evidence for the scientific community to reach defensible conclusions regarding the average effects of teacher evaluation on student achievement scores, less evidence about the conditions in which evaluation works and its cost-effectiveness, and no rigorous work focused on rural settings, which has left most local policymakers in the dark. We addressed (but did not close) these gaps in the literature by examining a rurally focused, university-created, fee-based teacher evaluation system, the Network for Educator Effectiveness (NEE). Furthermore, NEE is popular among those districts eligible to adopt it despite the availability of alternative teacher evaluation systems without fees, underscoring its relevance.

We first compared NEE's fees to other teacher evaluation systems and found that NEE's $3 per student per year membership fee is lower, possibly substantially lower than districts would otherwise pay to implement a comparable system. Like prior work concerning teacher evaluation costs (Chambers et al., 2013), we do not assert that NEE's fee captures all costs required to design, implement, and maintain a teacher evaluation system. Nevertheless, NEE's fee arguably represents the cost of interest to districts. Moreover, linking the effects of evaluation to its fees, despite their limited information about indirect costs, is an improvement over prior work that only reports the effects. NEE's meager fees imply that even small effects may be worth the expenditure relative to other alternatives.

However, on average, NEE does not affect rural student math or reading achievement. All point estimates from our full samples (Table 3) are statistically insignificant, near-zero, and negligible according to work regarding effect size interpretation (Jacob et al., 2019; Kraft, 2020). Our conclusion regarding average effects is consistent with prior work from urban and national settings, which finds no math or reading effects or an effect in one subject only (Bleiberg et al., 2024; Steinberg & Sartain, 2015; Taylor & Tyler, 2012). While the absence of average effects suggests that policymakers might not unconditionally adopt NEE or systems like it, this does not mean that NEE never improves rural student achievement. Indeed, this is why we pushed beyond average effects and examined the conditions in which NEE works.

NEE may improve reading achievement in rural schools where the average student's prior-year achievement score is below the state average or the average teacher's years of experience are below the state average; the evidence repeatedly suggests that NEE causally improves math achievement in these schools. Although reading achievement increased in NEE schools with below-average prior-year achievement or teacher years of experience, we could not rule out that these improvements occurred due to chance or confounders. However, the data never suggested threats to the internal validity of estimates regarding math scores; we conclude that those effects are plausibly causal.

Among rural schools with below-average prior-year student achievement, NEE increased math scores by approximately 0.07 SD or 2.3 months of learning and improved math scores by about 0.04 SD or 1.3 months of student learning in rural schools where the typical teacher had below-average years of experience.[9] Importantly, these one-year-after effects are plausible and

[9] The average student can expect to gain 0.40 SD of learning, as measured by standardized test scores in one calendar year (Hill et al., 2008). Therefore, we approximate months of learning by dividing 0.40 by 12 (months), which is equal to 0.03 SD of learning per month.

representative of estimates from prior work in urban settings, which also focus on one-year-after effects (Steinberg & Sartain, 2015; Taylor & Tyler, 2012). Notably, NEE's effects-to-expenditure ratios in these below-average rural schools range from 0.013 SD to 0.023 SD per dollar spent per student. To place these ratios in context, Harter (1999) reports that increasing teacher salary supplements by $1 per teacher (in 2012 dollars) is associated with an increase in student math achievement scores of 0.0006 SD and Wenglinsky (1997) finds that increasing PPE assigned to the broad category of "instructional expenditures" by one 2012 dollar is associated with a rise of 0.000003 SD in mathematics.

Analyses regarding NEE's temporal effects on math scores in the below-average rural schools examined reveal context-dependent variability. In rural schools with below-average prior-year achievement, NEE improves math achievement one year after treatment, implying that NEE may be an enticing and cost-effective intervention for below-average schools facing immediate pressure to improve math scores. Simultaneously, realizing positive effects in schools with high concentrations of inexperienced teachers takes two years. Schools with large concentrations of inexperienced teachers must impart a great deal of professional knowledge to their staff. We speculate that NEE - which introduces new teacher performance expectations and measures and substantially increases the number of performance feedback episodes teachers receive - may initially overwhelm a staff already acquiring substantial professional knowledge. In such contexts, teachers may need additional time to incorporate all of this knowledge into practices affecting student math achievement. While we believe NEE can help schools with large concentrations of inexperienced teachers improve math achievement, leaders in these settings should not expect immediate benefits.

Our study examined school settings like those in Steinberg and Sartain's analysis of Chicago teacher evaluation (2015), yet they found Matthew effects. We conjecture that these diametric results stem from the value-add of a new developmentally oriented intervention, like Chicago's EITP and Missouri's NEE. If we assume that Chicago already targeted robust professional development to the schools needing the most improvement (i.e., those with below-average student achievement and less experienced teachers), adopting another developmentally focused intervention (EITP) may push these schools toward diminishing marginal returns. However, rural schools, which often receive less robust and frequent professional development than urban schools (Skyhar, 2020), may still be at a point where a new developmentally focused intervention (NEE) results in rising marginal returns. Future work might test the validity of this conjecture.

Limitations

This study may be limited in several ways. First, our findings may not generalize to other settings or to differently designed teacher evaluation systems. Second, NEE's effects on math (and potentially reading) scores in the below-average school settings examined may change over longer periods. Future work with longer panels might explore these effects in rural settings. Third, we only examine student achievement outcomes, but NEE may affect other student or educator outcomes. Indeed, we assume that NEE’s users believe it affects important unexamined outcomes positively; otherwise, we cannot fathom why districts would choose to join the fee-based NEE system. NEE’s growing popularity since the early 2010s bolsters our assumption as NEE has either been the most popular or second-most popular evaluation system adopted by rural Missouri districts and has expanded into rural Nebraska and Kansas. Finally, as discussed, we report effects-to-expenditure ratios, falling short of the ideal cost-effectiveness ratios.

Implications

Our work affords targeted policy implications, which we offer while urging caution befitting a single study. We do not advise rural districts to adopt systems like NEE unconditionally; instead, it may be better to think of these systems as interventions that can help rural schools with below-average prior-year student achievement and teacher years of experience improve math (and potentially) reading scores. Furthermore, we advise rural schools with high concentrations of inexperienced teachers to expect meaningful but delayed effects. Knowing NEE's fee can also help rurally minded policymakers choose the right intervention for improving achievement scores. The fee districts pay to NEE for math achievement score improvements between 0.04 and 0.07 SD suggests that NEE is incredibly cost-effective in these settings. Indeed, policymakers wanting to improve math scores in these contexts may be hard-pressed to find a more affordable alternative. State policymakers might incentivize rural schools with below-average prior-year student achievement and teacher years of experience to adopt NEE-like systems by assigning state-provided funds to these schools.

References

Almy, S. (2011). *Fair to Everyone: Building the Balanced Teacher Evaluations that Educators and Students Deserve* (Teacher Quality). The Education Trust. https://edtrust.org/resource/fair-to-everyone-building-the-balanced-teacher-evaluations-that-educators-and-students-deserve/

Bergin, C., Wind, S. A., Grajeda, S., & Tsai, C.-L. (2017). Teacher evaluation: Are principals' classroom observations accurate at the conclusion of training? *Studies in Educational Evaluation*, *55*, 19–26. https://doi.org/10.1016/j.stueduc.2017.05.002

Bleiberg, J., Brunner, E., Harbatkin, E., Kraft, M. A., & Springer, M. G. (2024). *Taking Teacher Evaluation to Scale: The Effect of State Reforms on Achievement and Attainment.* (Working Paper 21–496; EdWorkingPaper). Annenberg Institute at Brown University. https://www.edworkingpapers.com/ai21-496

Bowser, K. M. & Hunter, S. B. (2022). Using Administrative Data to Describe the Implementation of Teacher evaluation: An Exploratory Study Utilizing Missouri Data. Annual Meeting of the Association for Education Finance and Policy. Denver, CO.

Callaway, B., & Sant'Anna, P. H. C. (2021). Difference-in-Differences with multiple time periods. *Journal of Econometrics*, *225*(2), 200–230. https://doi.org/10.1016/j.jeconom.2020.12.001

Chambers, J., Brodziak de los Reyes, I., & O'Neil, C. (2013). *How Much Are Districts Spending to Implement Teacher Evaluation Systems? Case Studies of Hillsborough County Public Schools, Memphis City Schools, and Pittsburgh Public Schools* (Working Paper WR-989-BMGF; RAND Working Paper). RAND Corporation.

Cherasaro, T. L., Brodersen, R. M., Reale, M. L., & Yanoski, D. C. (2016). *Teachers' responses to feedback from evaluators: What feedback characteristics matter?* (REL 2017-190; Making Connections, pp. 1–29). REL Central.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). The Long Term Impacts of Teachers: Teacher Value Added and Student Outcomes in Adulthood. *American Economic Review*, *104*(9), 2633–2679.

Choi, E., & Johnson, D. A. (2022). Common Antecedent Strategies within Organizational Behavior Management: The Use of Goal Setting, Task Clarification, and Job Aids. *Journal of Organizational Behavior Management*, *42*(1), 75–95. https://doi.org/10.1080/01608061.2021.1967834

Cinelli, C., & Hazlett, C. (2020). Making Sense of Sensitivity: Extending Omitted Variable Bias. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *82*(1), 39–67. https://doi.org/10.1111/rssb.12348

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2006). Teacher-Student Matching and the Assessment of Teacher Effectiveness. *The Journal of Human Resources*, *41*(4), 778–820.

Cullen, J. B., Koedel, C., & Parsons, E. (2021). The Compositional Effect of Rigorous Teacher Evaluation on Workforce Quality. *Education Finance and Policy*, *16*(1), 7–41. https://doi.org/10.1162/edfp_a_00292

Dee, T. S., & Wyckoff, J. (2015). Incentives, Selection, and Teacher Performance: Evidence from IMPACT. *Journal of Policy Analysis and Management*, *34*(2). https://doi.org/10.1002/pam

Donaldson, M. L. (2021). *Multidisciplinary Perspectives on Teacher Evaluation: Understanding the Research and Theory* (1st ed.). Routledge.

Donaldson, M. L., & Woulfin, S. (2018). From Tinkering to Going “Rogue”: How Principals Use Agency When Enacting New Teacher Evaluation Systems. *Educational Evaluation and Policy Analysis*, *40*(4), 531–556. https://doi.org/10.3102/0162373718784205

Feng, L. (2010). Hire Today, Gone Tomorrow: New Teacher Classroom Assignments and Teacher Mobility. *Education Finance and Policy*, *5*(3), 278–316. https://doi.org/10.1162/EDFP_a_00002

Gilles, J. F. (2017). " It's Not a Gotcha": Interpreting Teacher Evaluation Policy in Rural School District. *The Rural Educator*, *38*(2).

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, *225*(2), 254–277. https://doi.org/10.1016/j.jeconom.2021.03.014

Gormley, T. A., & Matsa, D. A. (2011). Growing Out of Trouble? Corporate Responses to Liability Risk. *Review of Financial Studies*, *24*(8), 2781–2821. https://doi.org/10.1093/rfs/hhr011

Harter, E. A. (1999). How Educational Expenditures Relate to Student Achievement: Insights from Texas Elementary Schools. *Journal of Education Finance*, *24*(3), 281–302.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child development perspectives*, *2*(3), 172-177.

Hunter, S. B. (2020). The Unintended Effects of Policy-Assigned Teacher Observations: Examining the Validity of Observation Scores. *AERA Open*, *6*(2). https://doi.org/10.1177/2332858420929276

Hunter, S. B., & Ege, A. (2021). Linking Student Outcomes to School Administrator Discretion in the Implementation of Teacher Observations. *Educational Administration Quarterly*, *57*(4), 607–640. https://doi.org/10.1177/0013161X211003134

Hunter, S. B., & Kho, A. (Online). The Effects of Teacher Evaluation Policy on Student Achievement and Teacher Turnover: Leveraging Teacher Accountability and Teacher Development. *Journal of Education Human Resources*. https://doi.org/10.3138/jehr-2023-0040

Hunter, S. B., Kho, A., & Bowser, K. (2023, March). *Policy-Assigned Teacher Observations, Their Implementation, and Student Discipline Outcomes: Main, Mediated, and Moderated Relationships*. Annual Meeting of the Association of Education Finance and Policy, Denver, CO.

Hunter, S. B., & Springer, M. G. (2022). Performance Feedback, Human Capital, and Teacher Performance: A Mixed-Methods Analysis. *Educational Evaluation and Policy Analysis*, *44*(3), 380–403. https://doi.org/10.3102/01623737211062913

Hunter, S. B., & Steinberg, M. P. (2022). *Do You Observe What I Observe? The Predictors and Consequences of Discordance in Teacher and Evaluator Ratings of Teacher Performance*. https://doi.org/10.26300/97K9-BR18

Ilgen, D. R., Fisher, C. D., & Taylor, M. S. (1979). Consequences of Individual Feedback on Behavior in Organizations. *Journal of Applied Psychology*, *64*(4), 349–371. https://doi.org/10.1037/0021-9010.64.4.349

Jackson, C. K. (2018). What Do Test Scores Miss? The Importance of Teacher Effects on Non–Test Score Outcomes. *Journal of Political Economy*, *126*(5), 2072–2107.

Jacob, R. T., Doolittle, F., Kemple, J., & Somers, M.-A. (2019). A Framework for Learning From Null Results. *Educational Researcher*, *48*(9), 580–589. https://doi.org/10.3102/0013189X19891955

Kalogrides, D., & Loeb, S. (2013). Different Teachers, Different Peers: The Magnitude of Student Sorting Within Schools. *Educational Researcher*, *42*(6), 304–316. https://doi.org/10.3102/0013189X13495087

Kimball, S. M., & Milanowski, A. (2009). Examining Teacher Evaluation Validity and Leadership Decision Making Within a Standards-Based Evaluation System. *Educational Administration Quarterly*, *45*(1).

Kraft, M. A. (2019). Teacher Effects on Complex Cognitive Skills and Social-Emotional Competencies. *Journal of Human Resources*, *54*(1), 1–36. https://doi.org/10.3368/jhr.54.1.0916.8265R3

Kraft, M. A. (2020). Interpreting Effect Sizes of Education Interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

Kraft, M. A., & Gilmour, A. F. (2016a). Can Principals Promote Teacher Development as Evaluators? A Case Study of Principals' Views and Experiences. *Educational Administration Quarterly*, *52*(5), 711–753. https://doi.org/10.1177/0013161X16653445

Kraft, M. A., & Gilmour, A. F. (2016b). Revisiting the Widget Effect: Teacher Evaluation Reforms and the Distribution of Teacher Effectiveness. *Association of Education Finance and Policy*, 1–31.

Liebowitz, D. D., Porter, L., & Bragg, D. (2022). The Effects of Higher-Stakes Teacher Evaluation on Office Disciplinary Referrals. *Journal of Research on Educational Effectiveness*, *15*(3), 475–509. https://doi.org/10.1080/19345747.2021.2015496

Liu, J., & Loeb, S. (2021). Engaging Teachers: Measuring the Impact of Teachers on Student Attendance in Secondary School. *Journal of Human Resources*, *56*(2), 343–379. https://doi.org/10.3368/jhr.56.2.1216-8430R3

Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist*, *57*(9), 705–717. https://doi.org/10.1037/0003-066X.57.9.705

Marsh, J. A., Bush-Mecenas, S., Strunk, K. O., Lincove, J. A., & Huguet, A. (2017). Evaluating Teachers in the Big Easy: How Organizational Context Shapes Policy Responses in New Orleans. *Educational Evaluation and Policy Analysis*, *39*(4), 539–570. https://doi.org/10.3102/0162373717698221

Murphy, K. R., & Cleveland, J. N. (1995). *Understanding Performance Appraisal: Social, Organizational, and Goal-Based Perspectives*. Sage Publications.

Nakajima, N. (2022). *Evidence-Based Decisions and Education Policymakers* [Dissertation]. Harvard University.

National Council on Teacher Quality. (2019). *NCTQ: Yearbook: State Teacher Policy Database*. National Council on Teacher Quality (NCTQ). https://www.nctq.org/yearbook/home

Nguyen, T. D., (2020). Examining the teacher labor market in different rural contexts: Variations by urbanicity and rural states. *AERA Open, 6*(4).

Nguyen, T.D., Pham, L. D., Crouch, M., & Springer, M.G. (2020) The correlates of teacher turnover: An updated and expanded meta-analysis of the literature. *Educational Research Review*, 31.

Papay, J. P. (2012). Refocusing the Debate: Assessing the Purposes and Tools of Teacher Evaluation. *Harvard Educational Review*, *82*(1), 123–141. https://doi.org/10.17763/haer.82.1.v40p0833345w6384

Phipps, A. (2022). Does Monitoring Change Teacher Pedagogy and Student Outcomes? *EdWorkingPaper*, *22–510*. https://doi.org/Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/7021-1x97

Phipps, A., & Wiseman, E. A. (2021). Enacting the Rubric: Teacher Improvements in Windows of High-Stakes Observation. *Education Finance and Policy*, *16*(2), 283–312. https://doi.org/10.1162/edfp_a_00295

Rigby, J. G. (2015). Principals' Sensemaking and Enactment of Teacher Evaluation. *Journal of Educational Administration*, *53*(3), 374–392. https://doi.org/10.1108/JEA-04-2014-0051

Rodriguez, L. A., Swain, W. A., & Springer, M. G. (2020). Sorting Through Performance Evaluations: The Influence of Performance Evaluation Reform on Teacher Attrition and Mobility. *American Educational Research Journal*, 000283122091098. https://doi.org/10.3102/0002831220910989

Rosenbaum, P. R., & Rubin, D. B. (1983). Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome. *Journal of the Royal Statistical Society Series B*, *45*(2), 212–218.

Schwartz, N., Kang, H., Loeb, S., Grissom, J., Bartanen, B., Cheatham, J., Chi, O., Donaldson, M., Lemos, R. F., Mellon, G., Moffitt, S., Nurshatayeva, A., Owens, J., Pinker, E., White, R., & Zimmerman, S. (2023). *STUDYING THE SUPERINTENDENCY: A CALL FOR RESEARCH*. Annenberg Institute at Brown University. https://annenberg.brown.edu/sites/default/files/Studying%20the%20Superintendency%20-%20Call%20for%20Research.pdf

Skyhar, C. (2020). Thinking Outside the Box: Providing Effective Professional Development for Rural Teachers. *Theory & Practice in Rural Education*, *10*(1), 42–72. https://doi.org/10.3776/tpre.2020.v10n1p42-72

Song, M., Wayne, A. J., Garet, M. S., Brown, S., & Rickles, J. (2021). Impact of Providing Teachers and Principals with Performance Feedback on Their Practice and Student Achievement: Evidence from a Large-Scale Randomized Experiment. *Journal of Research on Educational Effectiveness*, 1–26. https://doi.org/10.1080/19345747.2020.1868030

Stecher, B. M., Garet, M. S., Hamilton, L. S., Steiner, E. D., Robyn, A., Poirier, J., Holtzman, D., Fulbeck, E. S., Chambers, J., & Brodziak de los Reyes, I. (2016). *Improving Teaching Effectiveness* (9780833092212). http://proxy.library.vcu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,url,cookie,uid&db=ehh&AN=11254922&site=ehost-live&scope=site

Steinberg, M. P., & Donaldson, M. L. (2016). The New Educational Accountability: Understanding the Landscape of Teacher Evaluation in the Post-NCLB Era. *Education Finance and Policy*, *11*(3). https://doi.org/10.1162/EDFP_a_00186

Steinberg, M. P., & Sartain, L. (2015). Does teacher evaluation improve school performance? Experimental evidence from Chicago's Excellence in Teaching Project. *Education Finance and Policy*, *10*(4), 535–572. https://doi.org/10.1162/EDFP_a_00173

Taylor, E. S., & Tyler, J. H. (2012). The Effect of Evaluation on Teacher Performance. *American Economic Review*, *102*(7), 3628–3651. https://doi.org/10.1257/aer.102.7.3628

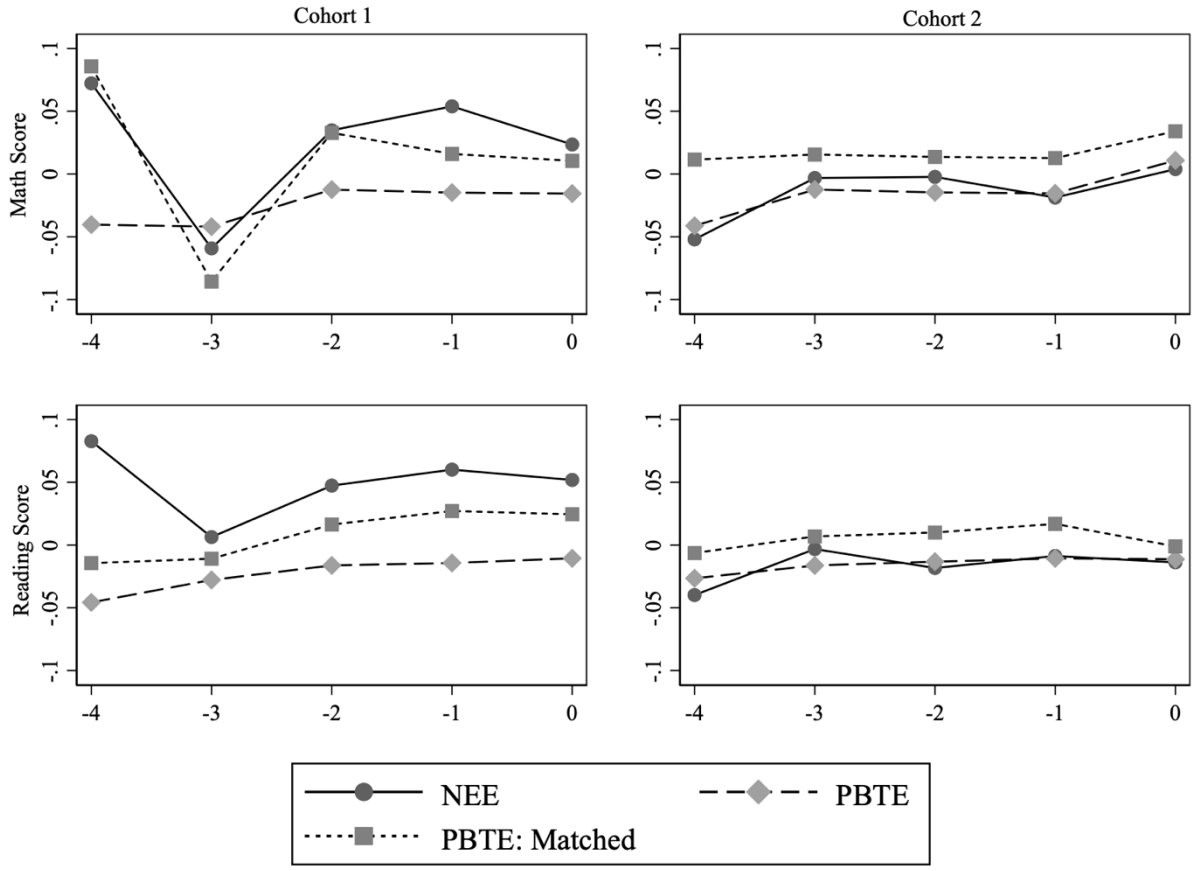
The School Superintendents Association. (2017). *Leveling the Playing Field for Rural Students*. https://files.eric.ed.gov/fulltext/ED594363.pdf

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The Widget Effect* (pp. 48–48). http://tntp.org/assets/documents/TheWidgetEffect_2nd_ed.pdf

Wenglinsky, H. (1997). School District Expenditures, School Resources and Student Achievement: Modeling the Production Function. In W. J. Jr. Fowler (Ed.), *Developments in School Finance 1997* (p. 196). National Center for Education Statistics.

Wind, S. A., Tsai, C.-L., Grajeda, S. B., & Bergin, C. (2018). Principals' use of rating scale categories in classroom observations for teacher evaluation. *School Effectiveness and School Improvement*, *29*(3), 485–510. https://doi.org/10.1080/09243453.2018.1470989

Figure 1. Average District-Level Student Achievement Scores Before and After NEE's Introduction



Notes: Each point represents average district-level achievement scores; districts are the unit of analysis. Year 0 represents NEE's introduction. The top panels plot math scores, the bottom panels plot reading scores, the left panels plot Cohort 1 trends, and the right panels Cohort 2 trends.

Table 1. CBO Ongoing Yearly and Total Per Pupil Expenditure Estimates

	Hillsborough County Public Schools	Memphis County Schools	Pittsburgh Public Schools
Year 1	$0.21 – $0.21	$0.42 – $0.42	$0 – $0
Year 2	$2.65 – $22.7	$3.34 – $3.66	$11.27 – $11.27
Year 3	$5.84 – $49.97	$26.86 – $37.73	$20.91 – $20.91
Total	$8.7 – $77.88	$30.62 – $41.81	$32.18 – $32.18

Notes. All costs per pupil dollars are adjusted to 2012 dollars. Ranges are conservative to liberal estimates based on disaggregated costs reported by Chambers et al. (2013).

Table 2. Descriptive Statistics

	NEE	Matched and Unmatched Non-NEE
Panel A. Student-Level Characteristics		
Prior-Year Math Score	0.01 (0.93) [16209]	0.01 (0.99) [470928]
Prior-Year Reading Score	0.02 (0.94) [16231]	0.01 (0.99) [474234]
Nonwhite	0.11 (.) [20535]	0.22 (.) [595834]
FRPL	0.54 (.) [20535]	0.50 (.) [595878]
Panel B. School-Level Characteristics		
School-Level Concentration Teacher More than MA	0.03 (.) [119]	0.03 (.) [4288]
School-Level Average Teacher Years of Experience	12.94 (2.33) [119]	12.82 (3.31) [4288]
Panel C. District-Level Characteristics		
Per Pupil Expenditure	8321.49 (1060.32) [30]	9969.60 (9498.87) [51069]
Rural	1.00 (.) [30]	0.84 (.) [1076]

Notes: Means, standard deviations in parentheses, and sample size in brackets. Descriptive statistics based on 2011-12 and 2012-13 records from NEE and non-NEE districts, matched or otherwise. Students are units of analysis in Panel A; schools are units in Panel B and districts in Panel C.

Table 3. NEE's Effect on Student Scores: Generalized Difference-in-Differences

	I	II	III	IV	V
Panel A. Math					
NEE	0.01	0.01	0.01	0.01	0.01
	(-0.02,0.05)	(-0.05,0.07)	(-0.10, 0.13)	(-0.05, 0.07)	(-0.05, 0.07)
N(Student-Yr)	319602	319602	319602	319602	319602
Panel B. Reading					
NEE	0.01	0.01	0.01	0.01	0.01
	(-0.00,0.03)	(-0.04, 0.06)	(-0.10, 0.12)	(-0.02, 0.05)	(-0.04, 0.06)
N(Student-Yr)	456232	456232	456232	456232	456232
Controls		X			X
District FE				X	X
Year FE				X	X
Cohort FE				X	X
Controls-Cohort			X		
Dist-Cohort FE	X	X	X		
Year-Cohort FE	X	X	X		

Notes: Point estimates and 95 percent confidence intervals in parentheses represent NEE's effect on student achievement scores. All models control for urbanicity, student prior-year math score, and district-level prior-year PPE. Standard errors are multiway clustered by district, student, and cohort. * $p < 0.05$

Table 4. School Characteristics Moderating NEE's Effects on Achievement Scores

	I	II	III	IV	V	VI
Panel A. Math						
NEE* Prior-Year Achievement < 0	0.04** (0.01,0.07)	0.06* (0.01,0.12)	0.07** (0.02,0.11)			
NEE* Prior-Year Achievement ≥ 0	-0.00 (-0.02,0.02)					
NEE*Avg Tch Yrs Exp < 12				0.04** (0.02,0.07)	0.05* (0.01,0.09)	0.04* (0.01,0.07)
NEE*Avg Tch Yrs Exp ≥ 12				0.01 (-0.02,0.04)		
N(Student-Yr)	319602	119927	119927	319602	101179	101179
Panel B. Reading						
NEE* Prior-Year Achievement < 0	0.03* (0.00,0.06)	0.03* (0.01,0.05)	0.03 (-0.06,0.13)			
NEE* Prior-Year Achievement ≥ 0	0.00 (-0.12,0.12)					
NEE*Avg Tch Yrs Exp < 12				0.02* (0.00,0.03)	0.03 (-0.06,0.12)	0.02 (-0.08,0.14)
NEE*Avg Tch Yrs Exp ≥ 12				-0.03 (-0.16,0.09)		
N(Student-Yr)	456232	193228	193228	456232	211135	211135
Controls			X			X

Notes: Models in columns I and IV applied to the full-matched sample and interact treatment with a moderator. Models in columns II and V applied to subsamples. Models in columns III and VI control for student and school observables. All models apply district-cohort fixed effects, year-cohort fixed effects, and control for urbanicity, student prior-year achievement score, and district-level prior-year PPE. Point estimates and 95 percent confidence intervals represent NEE's total effects on student achievement. Standard errors are multiway clustered by district, student, and cohort. * $p < 0.05$, ** $p < 0.01$.

Table 5. Nonparametric Event Study Estimates for Content-Specific Subsamples

	Math	Reading
	I	II
Panel A. School Prior-Year Achievement < 0 Sample		
4 Yrs Before NEE x Treated	-0.01	-0.02
	(0.01)	(0.01)
3 Yrs Before NEE x Treated	-0.01	0.01
	(0.02)	(0.00)
2 Yrs Before NEE x Treated	0.00	0.02
	(0.01)	(0.01)
NEE's First Yr x Treated	0.07*	0.03*
	(0.01)	(0.01)
N(Student-Yr)	119927	193228
Panel B. Average Teacher Years of Experience < 12 Sample		
4 Yrs Before NEE x Treated	0.01	-0.00
	(0.01)	(0.01)
3 Yrs Before NEE x Treated	-0.00	-0.00
	(0.04)	(0.02)
2 Yrs Before NEE x Treated	0.02	0.00
	(0.01)	(0.01)
NEE's First Yr x Treated	0.05*	0.01
	(0.01)	(0.02)
N(Student-Yr)	101179	211135

Notes: Point estimates and 95 percent confidence intervals. Estimates represent NEE's 'effects' on achievement scores relative to the 'effect' one year before NEE's introduction. Students are the unit of analysis. Models apply district-cohort fixed effects, year-cohort fixed effects, and controls for urbanicity, student prior-year achievement score, and district-level prior-year PPE. Standard errors are multiway clustered by district, student, and cohort. The sample in Panel A is restricted to the subsample of schools where the average student's achievement score was below zero. The sample in Panel B is restricted to the subsample of schools where the average teacher's years of experience is below the state average. * $p < 0.05$.

Table 6. Content-Specific Subsample Effects: Within and Between Cohort Effects

	Math			Reading		
	I	II	III	IV	V	VI
Panel A. School Prior-Year Achievement < 0 Sample						
NEE Cohort 1	0.04			0.03*		
	(-0.06,0.15)			(0.01, 0.05)		
NEE Cohort 2	0.08*			0.03*		
	(0.06,0.10)			(0.01, 0.05)		
NEE Year 1		0.06**	0.09***		0.02	0.01
		(0.02,0.10)	(0.05,0.13)		(-0.00,0.04)	(-0.04,0.06)
NEE Year 2		0.03			0.02	
		(-0.01,0.07)			(-0.00,0.04)	
N(Student-Yr)	119927	43513	34197	193228	62660	31466
Panel B. Average Teacher Years of Experience < 12 Sample						
NEE Cohort 1	-0.03			-0.00		
	(-0.21,0.14)			(-0.02,0.02)		
NEE Cohort 2	0.06*			0.03		
	(0.00, 0.13)			(-0.02,0.08)		
NEE Year 1		-0.03	0.07***		-0.03	0.01
		(-0.08,0.01)	(0.03,0.11)		(-0.10,0.04)	(-0.02,0.05)
NEE Year 2		0.06*			0.00	
		(0.01,0.11)			(-0.09,0.09)	
N(Student-Yr)	101179	35898	24583	211135	75820	26090
Cohort 1	X	X	X	X	X	X
Cohort 2	X		X			X

Notes: Columns I and IV moderate NEE's effect by cohort. Columns II and V moderate the effects of Cohort 1 only by year. Columns III and VI apply Equation 1 to 2007-08 through 2011-12 using data from Cohorts 1 and 2; in 2011-12, Cohort 1 was in its first year of NEE implementation, and Cohort 2 was in its last year of non-NEE implementation. Point estimates and 95 percent confidence intervals represent total effects on student achievement scores. Standard errors in columns I, III, IV, and VI are multiway clustered by district, student, and cohort; standard errors in columns II and V are clustered by district and student. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 7. Sensitivity of Effects to Omitted Variables with Benchmarked Relationships for Content-Specific Subsamples

	Math			Reading		
	I	II	III	IV	V	VI
	$R^2_{NEE\sim OV\|X}$	$R^2_{Y\sim OV\|X,NEE}$	Coef	$R^2_{NEE\sim OV\|X}$	$R^2_{Y\sim OV\|X,NEE}$	Coef
Panel A. School Avg Prior-Year Achievement < 0						
0.1 x Prior-Year Ach	0.000	0.318	0.06*	0.000	0.290	0.04*
0.2 x Prior-Year Ach	0.000	0.636	0.06*	0.000	0.579	0.04*
0.3 x Prior-Year Ach	0.000	0.954	0.06*	0.000	0.869	0.03*
N(Student-Yr)		119927			193228	
Panel B. Average Teacher Years of Experience < 12						
0.1 x Prior-Year Ach	0.000	0.269	0.03*	0.000	0.268	0.03
0.2 x Prior-Year Ach	0.000	0.538	0.03*	0.000	0.536	0.03
0.3 x Prior-Year Ach	0.000	0.807	0.03*	0.000	0.804	0.02
N(Student-Yr)		101179			211135	

Notes: Models apply Equation 1. Standard errors from models in columns III and VI in Table 4. *NEE* represents treatment, *OV* the hypothetical omitted variables, and *X* all righthand-side variables from Equation 1 excluding *NEE*. $R^2_{NEE\sim OV|X}$ represents the residual variation in *NEE*. $R^2_{Y\sim OV|X,NEE}$ represents the residual variation in student achievement scores. "Coef" is the estimated treatment effect if Equation 1 controlled for *OV*. * $p < 0.05$

Online Appendix A. Effects on Covariates

We estimate effects on covariates using Equation A:

$$x_{isdt} = \delta NEE_{dt} + \beta_1 y_{isd(t-1)} + \beta_2 PPE_{d(t-1)} + \Delta_{dc} + \Phi_{tc} + e_{isdtc} \quad (A),$$

where x_{isdt} represents student-level gender, nonwhite racial status, FRPL, and prior-year achievement, and the school-level and district-level concentrations of female, nonwhite, and FRPL students, and the average student's prior-year achievement score. A statistically significant δ in Equation A would imply that NEE affects outcomes it should not affect, theoretically, undermining a causal interpretation of the results from Equation 1.

Online Appendix B. NEE Cost Analyses

Table B1. CBO and NEE Alignment

CBO Category	CBO Subdomain	NEE Alignment Rationale
Design and Implementation	*…expenditures associated with the **development of materials and processes**…such as the **teacher observation rubric** and the VAM. Activities associated with implementation, such as **trainings and observer calibrations**… (p. 7)*	NEE provides already developed materials including an **observation rubric.** NEE staff provide **annual and ongoing training** to NEE implementers, including required **observer calibrations**.
Peer, mentor, and external evaluators	*…includes the **salaries, benefits, and travel expenses of full-time peer and mentor observers**…Though school leaders conduct observations in all three districts, we are not able to account for their time spent on initiative-related activities in our expenditure estimates.* (p.8)	N/A
Management and communications	*…relate to activities regarding the **planning and implementation** of the effective teaching initiatives and the **communication efforts to introduce the reforms to the district staff** within each of the three teacher evaluation components.* (p. 8)	NEE is heavily involved in NEE rollout. NEE's plethora of **ready-made resources, including language and training materials, and accessibility to NEE staff support district communications efforts**, which otherwise would be more time and cost intensive.
Technology and data systems	*…investments that the districts made **to develop software infrastructure as well as to purchase information technology (IT) equipment** to support the teacher observation, VAM, and survey components.* (p. 8)	NEE provides all NEE users with access to the NEE Data Tool, a centrally managed **data management system**. Further, NEE staff provide all **technical support** via the NEE Help Desk.
Other	*…such as office **overhead and support for data collection**.* (p. 8)	NEE provides **support that may be found in the central offices of larger districts** (e.g., training, data management). Due to the vagueness of this category, we include in the upper bound cost estimates.

Notes. CBO subdomain definitions are quotes from Chambers et al. (2013).

Table B2. CBO Start-Up Yearly and Total Per Pupil Expenditure Estimates

	Hillsborough County Public Schools	Memphis County Schools	Pittsburgh Public Schools
Year 1	$11.46 – $11.46	$5.44 – $5.64	$27 – $27
Year 2	$28.22 – $28.22	$14.11 – $16.2	$31.78 – $31.78
Year 3	$6.58 – $6.58	$4.7 – $4.7	$29.34 – $29.34
Total	$46.26 – $46.26	$30.62 – $41.81	$88.12 – $88.12

Notes. All costs per pupil dollars are adjusted to 2012 dollars. Ranges are conservative to liberal estimates based on disaggregated costs reported by Chambers et al. (2013).

Online Appendix C. Coarsened Exact Matching Results

Table C1. Math Sample Matched Results

	Cohort 1		Cohort 2	
	L1	Mean	L1	Mean
District-level average student math achievement scores				
t = 2006-07	0.36	-0.03	0.19	-0.01
t = 2007-08	0.27	-0.01	0.18	-0.00
t = 2008-09	0.07	0.01	0.15	-0.00
t = 2009-10	0.29	0.00	0.16	-0.00
t = 2010-11	0.36	0.03	0.16	0.00
t = 2011-12			0.22	-0.03
District-level PPE				
t = 2006-07	0.26	$100.52	0.23	$2.60
t = 2007-08	0.48	$161.02	0.28	-$247.78
t = 2008-09	0.54	$129.05	0.21	-$72.98
t = 2009-10	0.42	$195.25	0.19	-$174.74
t = 2010-11	0.41	- $24.46	0.21	-$54.46
t = 2011-12			0.18	$129.99
Urbanicity	0.00	0.00	0.00	0.00

Notes: Districts are the unit of analysis. The multivariate L1 distance for Cohorts 1 and 2 is 1.0. Cohort 1 data from 2012 are purposefully omitted as outcomes for this cohort are measured in 2012. The sample of potential matches for Cohort 1 in 2010-11 included districts that would join NEE in 2011-12 but had not yet in 2010-11. Cohort 1 is always excluded from Cohort 2's potential matches.

Table C2. Reading Sample Matched Results

	Cohort 1		Cohort 2	
	L1	Mean	L1	Mean
District-level average student reading achievement scores				
t = 2006-07	0.46	-0.00	0.16	-0.03
t = 2007-08	0.68	0.09	0.20	-0.03
t = 2008-09	0.20	0.01	0.11	-0.00
t = 2009-10	0.37	0.04	0.23	-0.00
t = 2010-11	0.52	0.04	0.22	-0.01
t = 2011-12			0.26	-0.01
District-level PPE				
t = 2006-07	0.25	$36.90	0.17	-$188.50
t = 2007-08	0.46	$65.92	0.19	-$385.57
t = 2008-09	0.33	$37.70	0.21	-$221.64
t = 2009-10	0.33	$114.77	0.23	-$262.91
t = 2010-11	0.48	- $17.69	0.24	-$242.42
t = 2011-12			0.28	-$111.84
Urbanicity	0.00	0.00	0.00	0.00

Notes: See Table C1 notes.

Online Appendix D. School Characteristics Moderating NEE's Effects on Achievement Scores

Table D1. Economic Disadvantage and Nonwhite Racial Proportions

	Math Scores		Reading Scores	
	I	II	III	IV
NEE* School FRPL < 50%	0.03		0.02	
	(-0.07,0.12)		(-0.04,0.08)	
NEE* School FRPL ≥ 50%	0.01		0.01	
	(-0.18,0.20)		(-0.04,0.06)	
NEE*School Proportion Nonwhite Students < 10%		0.01		0.02
		(-0.02,0.04)		(-0.01,0.04)
NEE*School Proportion Nonwhite Students ≥ 10%		0.01		0.01
		(-0.02,0.05)		(-0.04,0.05)
N(Student-Yr)	319602	319602	456232	456232

Notes: Models in columns I and III interact treatment with an indicator representing whether the school's proportion of FRPL students is below 50% or not. Models in columns II and IV interact treatment with an indicator representing whether the school's proportion of nonwhite students is below 10% or not. All models apply district-cohort fixed effects, year-cohort fixed effects, and control for urbanicity, student prior-year achievement score, district-level prior-year PPE, and the binary moderator. Point estimates and 95 percent confidence intervals represent NEE's total effects on student achievement scores within the subgroups. Standard errors multiway clustered by district, student, and cohort.

Online Appendix E. Sensitivity Analyses

Table E1. Nonparametric Event Study Estimates for Full Sample

	Math	Reading
	I	II
4 Yrs Before NEE x Treated	0.01	0.00
	(-0.40,0.42)	(-0.17,0.18)
3 Yrs Before NEE x Treated	0.03*	0.01
	(0.00,0.06)	(-0.36,0.39)
2 Yrs Before NEE x Treated	0.02	0.00
	(-0.18,0.23)	(-0.25,0.26)
NEE's First Yr x Treated	0.03	0.01
	(-0.18,0.23)	(-0.12,0.14)
N(Student-Yr)	319602	456232

Notes: Point estimates and 95 percent confidence intervals. Estimates represent NEE's 'effects' on achievement scores relative to the 'effect' one year prior to NEE's introduction. Students are the unit of analysis. Models apply district-cohort fixed effects, year-cohort fixed effects, and controls for urbanicity, student prior-year achievement score, and district-level prior-year PPE. Standard errors multiway clustered by district, student, and cohort. * $p < 0.05$, ** $p < 0.01$.

Table E2. NEE's Effect: Cohort 1 vs Cohort 2 Sample and Cohort-Specific

	Math		Reading	
	I	II	III	IV
NEE: Cohort 1	0.01	0.02	0.02	0.01
	(-0.04, 0.06)	(-0.01, 0.05)	(-0.05, 0.09)	(-0.04, 0.06)
NEE: Cohort 2	0.02		0.01	
	(-0.01, 0.04)		(-0.03, 0.05)	
N(Student-Yr)	319096	78885	456232	83744
Matched Sample	X		X	
C1vsC2 Sample		X		X

Notes: All models control for urbanicity, student prior-year math score, and district-level prior-year PPE. Models I and III apply district-cohort and year-cohort fixed effects. Columns II and IV apply Equation 1 to 2007-08 through 2011-12 using data from Cohorts 1 and 2 only; in 2011-12, Cohort 1 was in its first year of NEE implementation, and Cohort 2 had not yet implemented NEE. Point estimates and 95 percent confidence intervals represent NEE's total effects. Standard errors in Models I and III multiway clustered by district, student and cohort; Models II and IV multiway clustered errors by district and student.

Table E3. Placebo Tests for Content-Specific Subsamples

	I	II	III	IV
Years Preceding NEE	t-1	t-2	t-3	t-4
Panel A. Math				
Panel A1. School Prior-Year Achievement < 0 Sample				
NEE	0.00	0.02	-0.01	0.00
	(-0.06, 0.06)	(-0.15, 0.18)	(-0.28, 0.26)	(-0.05, 0.05)
N(Student-Yr)	119927	119927	119927	119927
Panel A2. Average Teacher Years of Experience < 12 Sample				
NEE	-0.01	0.00	-0.04	0.00
	(-0.21, 0.18)	(-0.15, 0.15)	(-0.51, 0.43)	(-0.52, 0.52)
N(Student-Yr)	101179	101179	101179	101179
Panel B. Reading				
Panel B1. School Prior-Year Achievement < 0 Sample				
NEE	-0.05*	0.01	-0.01*	0.02*
	(-0.07, -0.04)	(-0.05, 0.07)	(-0.02, -0.00)	(0.01, 0.03)
N(Student-Yr)	193228	193228	193228	193228
Panel B2. Average Teacher Years of Experience < 12 Sample				
NEE	-0.01	0.02	0.02	-0.03
	(-0.18, 0.16)	(-0.32, 0.36)	(-0.16, 0.21)	(-0.07, 0.02)
N(Student-Yr)	211135	211135	211135	211135

Notes: Point estimates and 95 percent confidence intervals in parentheses represent NEE's 'effect' on achievement scores in years preceding NEE's introduction. Models apply district-cohort fixed effects, year-cohort fixed effects, and control for urbanicity, student prior-year achievement score, and district-level prior-year PPE. Panels A1 and B1 models applied to the subsample of schools where the average student's achievement score was below zero. Panels A2 and B2 models were applied to the subsample of schools where the average teacher's years of experience were below the state average. Standard errors are multiway clustered by district, student, and cohort. * $p < 0.05$.

Table E4. Placebo Tests

	I	II	III	IV
Years Preceding NEE	t-1	t-2	t-3	t-4
Panel A. Math Scores				
NEE	0.00	0.02	-0.01	0.00
	(-0.06,0.06)	(-0.15,0.18)	(-0.28,0.26)	(-0.05,0.05)
N(Student-Yr)	319602	319602	319602	319602
Panel B. Reading Scores				
NEE	0.00	-0.01	0.01	-0.00
	(-0.15, 0.15)	(-0.18, 0.18)	(-0.24, 0.26)	(-0.02, 0.01)
N(Student-Yr)	456232	456232	456232	456232

Notes: Point estimates and 95 percent confidence intervals in parentheses represent NEE's 'effect' on achievement scores in years preceding NEE's introduction. Models apply district-cohort fixed effects, year-cohort fixed effects, and control for urbanicity, student prior-year achievement score, and district-level prior-year PPE. Standard errors multiway clustered by district, student, and cohort.

Online Appendix F. Compositional Effects

Table F1. Effects on Observable Characteristics for Schools with Prior-Year Achievement < 0

Panel A. Student Characteristics	Math Students		Reading Students	
Female	0.00	(-0.10, 0.09)	0.00	(-0.01, 0.01)
Nonwhite	-0.00	(-0.04, 0.05)	-0.01	(-0.03, 0.01)
FRPL	-0.03	(-0.08, 0.03)	-0.01*	(-0.02, -0.01)
Prior-Year Achievement Score	-0.01	(-0.05, 0.03)	-0.01	(-0.02, 0.00)
Panel B. School Characteristics				
Concentration Female Students	0.00	(-0.03, 0.04)	-0.01	(-0.02, 0.00)
Concentration Nonwhite Students	-0.00	(-0.06, 0.07)	-0.01	(-0.03, 0.01)
Concentration FRPL Students	-0.03	(-0.10, 0.05)	-0.01*	(-0.02, -0.00)
Avg Stdt Prior-Yr Ach Score	-0.02	(-0.09, 0.05)	-0.01	(-0.02, 0.00)
Concentration Female Teachers	0.01	(-0.15, 0.18)	0.00	(-0.02, 0.03)
Concentration Nonwhite Teachers	0.00	(-0.01, 0.01)	0.00*	(0.00, 0.01)
Concentration Adv Degrees	0.00	(-0.09, 0.08)	0.00	(-0.00, 0.01)
Avg Tch Years of Experience	-0.47	(-7.66, 6.72)	-0.20	(-0.85, 0.45)
Panel C. District Characteristics				
Concentration Female Students	0.00	(-0.01, 0.01)	-0.01	(-0.02, 0.00)
Concentration Nonwhite Students	0.02	(-0.01, 0.05)	0.01	(-0.01, 0.03)
Concentration FRPL Students	-0.02	(-0.05, 0.01)	-0.01	(-0.03, 0.00)
Avg Stdt Prior-Yr Ach Score	-0.01	(-0.04, 0.02)	-0.01	(-0.02, 0.00)
Concentration Female Teachers	0.01	(-0.01, 0.02)	0.01	(-0.00, 0.03)
Concentration Nonwhite Teachers	-0.00	(-0.00, 0.00)	0.00	(-0.00, 0.00)
Concentration Adv Degrees	0.01	(-0.03, 0.05)	0.01*	(0.00, 0.02)
Avg Tch Years of Experience	0.29	(-0.07, 0.66)	0.09	(-0.33, 0.51)
Per Pupil Expenditure	-157.94	(-2396.33, 2080.45)	37.36	(-3468.37, 3543.09)
Prior-Year Per Pupil Expenditure	-95.45	(-337.03, 146.12)	9.41	(-0.28, 1.11)
N(Student-Yr)	119927		193228	

Notes: Point estimates and 95 percent confidence intervals in parentheses represent NEE's 'effect' on each covariate. A different regression generates each row. Models apply district-cohort fixed effects, year-cohort fixed effects, and control for urbanicity, student prior-year achievement score, and district-level prior-year PPE. Standard errors are multiway clustered by district, student, and cohort. * $p < 0.05$

Table F2. Effects on Observable Characteristics for Schools with Avg Tch Experience < 12

Panel A. Student Characteristics	Math Students		Reading Students	
Female	-0.01	(-0.05, 0.03)	0.00	(-0.03, 0.03)
Nonwhite	-0.00	(-0.03, 0.04)	0.00	(-0.03, 0.03)
FRPL	-0.02	(-0.14, 0.10)	-0.01	(-0.09, 0.07)
Prior-Year Achievement Score	-0.08	(-0.23, 0.06)	-0.05	(-0.17, 0.07)
Panel B. School Characteristics				
Concentration Female Students	0.00	(-0.00, 0.00)	0.00	(-0.02, 0.01)
Concentration Nonwhite Students	-0.00	(-0.02, 0.02)	0.00	(-0.02, 0.01)
Concentration FRPL Students	-0.01	(-0.12, 0.10)	-0.01	(-0.10, 0.08)
Avg Stdt Prior-Yr Ach Score	-0.07	(-0.19, 0.04)	-0.05	(-0.17, 0.07)
Concentration Female Teachers	-0.01	(-0.11, 0.09)	0.04	(-0.13, 0.21)
Concentration Nonwhite Teachers	0.01	(-0.02, 0.03)	0.01	(-0.01, 0.02)
Concentration Adv Degrees	0.01	(-0.04, 0.06)	0.01	(-0.06, 0.07)
Avg Tch Years of Experience	0.66	(-0.64, 1.97)	-0.12	(-0.96, 0.72)
Panel C. District Characteristics				
Concentration Female Students	0.00	(-0.01, 0.00)	0.00	(-0.02, 0.01)
Concentration Nonwhite Students	-0.00	(-0.43, 0.42)	0.00	(-0.02, 0.01)
Concentration FRPL Students	-0.01	(-0.05, 0.04)	-0.01	(-0.04, 0.02)
Avg Stdt Prior-Yr Ach Score	-0.04	(-0.13, 0.06)	-0.02	(-0.13, 0.08)
Concentration Female Teachers	0.01	(-0.08, 0.10)	0.01	(-0.06, 0.08)
Concentration Nonwhite Teachers	-0.00	(-0.01, 0.01)	0.00	(-0.01, 0.01)
Concentration Adv Degrees	0.01	(-0.05, 0.06)	0.01	(-0.04, 0.05)
Avg Tch Years of Experience	0.51	(-0.55, 1.57)	0.01	(-0.56, 0.58)
Per Pupil Expenditure	17.36	(-1687.45, 1722.16)	-186.15	(-834.31, 462.02)
Prior-Year Per Pupil Expenditure	-215.48	(-632.89, 201.94)	167.78	(-3462.87, 3798.42)
N(Student-Yr)	101179		211112	

Notes: Point estimates and 95 percent confidence intervals in parentheses represent NEE's 'effect' on each covariate. A different regression generates each row. Models apply district-cohort fixed effects, year-cohort fixed effects, and control for urbanicity, student prior-year achievement score, and district-level prior-year PPE. Standard errors are multiway clustered by district, student, and cohort. * $p < 0.05$

Table F3. Compositional 'Effects' on Observable Characteristics

Panel A. Student Characteristics	Math Students		Reading Students	
Female	0.00	(-0.10, 0.10)	0.00	(-0.07,0.06)
Nonwhite	0.00	(-0.01, 0.01)	0.00	(-0.03,0.03)
FRPL	-0.01	(-0.14, 0.13)	0.00	(-0.06,0.06)
Panel B. School Characteristics				
Concentration Female Students	0.00	(-0.08, 0.08)	0.00	(-0.04,0.04)
Concentration Nonwhite Students	0.00	(-0.01, 0.01)	0.00	(-0.05,0.04)
Concentration FRPL Students	0.00	(-0.10, 0.10)	0.00	(-0.06,0.06)
Concentration Female Teachers	-0.01	(-0.08, 0.05)	0.00	(-0.09,0.08)
Concentration Nonwhite Teachers	0.00	(-0.04, 0.04)	0.00	(-0.03,0.03)
Concentration Adv Degrees	0.01	(-0.03, 0.04)	0.01	(-0.02,0.03)
Avg Tch Years of Experience	0.08	(-3.08, 3.24)	0.01	(-2.14,2.15)
Panel C. District Characteristics				
Concentration Female Students	0.00	(-0.07, 0.07)	0.00	(-0.03,0.03)
Concentration Nonwhite Students	0.00	(-0.34, 0.34)	0.00	(-0.03,0.03)
Concentration FRPL Students	-0.01	(-0.08, 0.07)	0.00	(-0.04,0.03)
Concentration Female Teachers	-0.01	(-0.10, 0.09)	0.00	(-0.14,0.14)
Concentration Nonwhite Teachers	0.00	(-0.02, 0.01)	0.00	(-0.01,0.01)
Concentration Adv Degrees	0.01	(-0.00, 0.02)	0.01	(-0.02,0.03)
Avg Tch Years of Experience	0.12	(-1.25, 1.50)	0.09	(-0.50,0.68)
Per Pupil Expenditure	-103.07	(-626.35, 420.22)	27.21	(-812.70, 867.13)
N(Student-Yr)	319602		456232	

Notes: Point estimates and 95 percent confidence intervals in parentheses represent NEE's effect on each post-treatment observable. Each row generated by a different regression. Models apply district-cohort fixed effects, year-cohort fixed effects, and control for urbanicity, student prior-year achievement score, and district-level prior-year PPE. Standard errors multiway clustered by district, student, and cohort.