Abstract

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), a non-profit evaluation system created for rural users. NEE does not affect student achievement on average but improves it substantially in disadvantaged rural schools; the positive effects-to-expenditure ratios in these settings are remarkable.

Keywords: evaluation, school/teacher effectiveness, educational policy, quasi-experimental analysis

Highlights

- Missouri districts are responsible for designing and maintaining reformed teacher evaluation systems or outsourcing these tasks to external organizations like the Network for Educator Effectiveness (NEE), a non-profit, university-based evaluation system designed for rural users, for a small fee.
- We compare student achievement trends in districts that adopted NEE to achievement in districts that did not. Critically, pre-NEE trends in adopting and non-adopting districts were comparable over four years prior to NEE's introduction. Therefore, we attribute deviations from these trends in NEE districts after adoption to NEE’s introduction.
- NEE does not affect rural student achievement on average but improves it substantially in disadvantaged rural schools. The effects-to-expenditure ratios in disadvantaged rural schools are remarkable compared to similar ratios in prior research, implying that education agencies might promote reformed teacher evaluation in these settings to improve schools.

Introduction

Incentivized by the historic Race to the Top competition, nearly every state has implemented a "next-generation" teacher evaluation system that includes standards-based observation rubrics, tenure reforms, and frequent, structured performance feedback conferences, among other features (Donaldson, 2021; National Council on Teacher Quality, 2019; Steinberg & Donaldson, 2016). According to state and local education agencies, these systems aim to improve teacher effectiveness via development (e.g., performance feedback) primarily and accountability (e.g., tenure reform) 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; J. Liu & Loeb, 2021). However, next-generation systems can be expensive (Stecher et al., 2016), and may impose substantial burdens on school administrators (Hunter & Rodriguez, 2021; 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 evaluation's effects on student achievement is relatively thin.[1] We have learned a great deal about teacher evaluation in a few urban centers (Dee & Wyckoff, 2015; Steinberg & Sartain, 2015; Taylor & Tyler, 2012) and one emerging national study (Bleiberg et al., 2021); these studies suggest that evaluations' effects on student achievement are mixed, at best. A smaller but important body of work has examined some of the conditions moderating these effects; one study finds that effects rise with teacher years of experience

[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., 2019). A multi-site randomized control trial also identifies the effect of providing educators with performance feedback, one aspect of next-generation evaluation, on student achievement scores (Song et al., 2021).

(Taylor & Tyler, 2012) while another concludes that evaluation's effects rise with school-level student economic advantage and prior-year achievement scores (Steinberg & Sartain, 2015). However, little work focuses on evaluations' effects on student achievement in rural settings, although most districts within most states are rural (National Center for Education Statistics, 2013). As emerging research finds education policymakers prioritize generalizability over internal validity (Nakajima, 2022), the absence of rigorous, rurally-situated teacher evaluation studies has left those who craft policy affecting rural schools in the dark. Moreover, Rodriguez and colleagues (2020) suggest that urbanicity might be a driver of evaluations' effects on teacher mobility, underscoring urbanicity's potential importance in understanding next-generation system effects. Additionally, we are unaware of any study with plausibly causal effects linking evaluation expenditures to effects. Ultimately, there is insufficient evidence for the scientific[2] community to reach defensible conclusions about next-generation evaluations' effects on rural student achievement scores, less evidence regarding the conditions in which evaluation improves student outcomes, and no rigorous research linking expenditures and effects.

This study's broad purpose is to advance our understanding of teacher evaluations' effects on student achievement by answering the following questions:

1. What are the effects of implementing a next-generation teacher evaluation system on student mathematics and reading achievement scores in rural settings?
2. To what extent do these effects vary by school-level teacher years of experience, student economic disadvantage, race, and prior-year achievement scores?

We estimate the impact of the Network for Educator Effectiveness (NEE), a fee-based Missouri next-generation teacher evaluation system designed and supported by a university-

[2] We purposefully apply the qualifier "scientific" as laypeople seem to have reached premature conclusions about teacher evaluation, as noted elsewhere (Cohen et al., 2020).

based organization within the University of Missouri, using a difference in differences framework applied to five years of panel data. Next-generation reforms often required the development of teacher performance measures (e.g., observation rubrics), evaluator training and supports, new evaluation procedures, and the development and implementation of other support systems (Archer et al., 2016; Chambers et al., 2013; Stecher et al., 2016). Thus, beginning in the early 2010s, Missouri's local education agencies faced a choice to meet evaluation reforms' expectations: develop and maintain systems on their own or outsource these tasks to external organizations. Missouri districts that chose NEE opted for the latter in exchange for a small fee. NEE was designed for rural districts specifically and eschews evaluation for accountability while emphasizing teacher development, consistent with many education agencies' conceptualizations of teacher evaluation (Almy, 2011). NEE's developmental focus may be more applicable to rural districts, which face small teacher labor supplies that may inhibit evaluation's accountability mechanism, as implied by Rodriguez and colleagues (2020). Thus, NEE is a broadly relevant system for analysis.

We find that NEE's introduction generated precisely estimated, negligibly positive, statistically insignificant main effects on student achievement, resembling findings from other settings. However, consistent with NEE's developmental aims, student achievement rose in disadvantaged rural schools substantially while teacher turnover was unaffected. Despite the mixed evidence regarding NEE's effects, the small fee of approximately $3 per student is money well spent in disadvantaged rural settings. Three dollars does not represent NEE's net or opportunity cost per student; it is the amount districts paid for NEE's services. Although we presume policymakers and academics prefer to know NEE's total net or opportunity cost, we also presume that they prefer to learn something about district expenditures over no cost information at all. With this mind, we find that NEE's effects-to-expenditure ratio in disadvantaged rural settings is remarkable compared to similar ratios (Harter, 1999; Wenglinsky, 1997).

This study makes three contributions to teacher evaluation research. It is the first to estimate plausibly causal effects of next-generation teacher evaluation on rural student achievement specifically. Second, it adds to the small body of causal evidence concerning the school conditions in which evaluation improves student achievement; that this study examines conditions in rural districts extends this contribution further. Third, it is the first to link reformed teacher evaluation expenditures to its effects.

Background

Next-Generation Teacher Evaluation Theory of Action

Theoretically, next-generation teacher evaluation systems improve teacher effectiveness through two mechanisms: a) teacher accountability that results in the forced or voluntary exit of ineffective teachers from the teacher workforce or b) teacher professional development that improves individual effectiveness (Donaldson, 2021; Papay, 2012; Phipps & Wiseman, 2021). The accountability mechanism operates through several sub-mechanisms. Next-generation systems include standards-based performance criteria and observation protocols mapping criteria onto performance levels (Steinberg & Donaldson, 2016). By including performance expectations in standards and protocols, these systems define teacher performance expectations. Moreover, the higher frequency of classroom observations and post-observation performance-feedback conferences characteristic of next-generation systems allow evaluators to clarify performance expectations for teachers (Donaldson, 2021; Hunter & Springer, 2021; Steinberg & Donaldson, 2016). Theoretically, 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; Weisberg et al., 2009); 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 (Phipps & Wiseman, 2021), ultimately raising student achievement as research links higher performance on standards-based observation protocols to higher student achievement (Daley & Kim, 2010; Kane et al., 2011).

The developmental components of next-generation evaluation reforms might also improve teaching quality independent of pure accountability mechanisms. Observation conferences can provide teachers with performance-enhancing strategies directly or indirectly. 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 professional development opportunities tailored to observation-identified area of weakness (e.g., targeted coaching; Donaldson, 2021), underscoring the importance of linkages between evaluation and professional development 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 (Hattie & Timperley, 2016; Hunter & Springer, 2021; Kimball & Milanowski, 2009).

Related Prior Studies

The literature review focuses on the causal effects of introducing next-generation teacher evaluation system on student achievement scores, which only a few studies examine.[3] In a

[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

unique randomized control trial, Steinberg and Sartain (2015) estimated the effects of a next-generation teacher evaluation pilot, the Excellence in Teaching Project (EITP). EITP, a low-stakes system without tenure or dismissal reforms, was implemented across two cohorts of elementary schools in Chicago Public Schools. While analyses of student math scores did not detect any effects, student reading scores increased significantly. Importantly, these effects were almost entirely concentrated in the first cohort of the pilot study. Cohort 2 schools, which did not receive the same level of administrative and implementation support as Cohort 1 schools, did not exhibit similar effects. This is the only study we know of that estimates the moderating 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 moderation by school-level shares of student race or average teacher years of experience.

A quasi-experimental study 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.

An emerging study using national data from the Stanford Education Data Archive estimated the effects of adopting evaluation reforms on math and reading achievement across states using an event study and difference-in-differences framework (Bleiberg et al., 2021). Unlike prior work, this study finds no effects on either math or reading achievement. The authors

---

performance feedback measures. As these treatments differ from the treatment of introducing a next-generation evaluation system, we do not discuss them further.

also examine heterogeneity by a) rigor of the evaluation system design and b) student characteristics within district-grade-year cells; there is little evidence of heterogeneous effects.

Evaluation for Teacher Development

Teacher professional development research also implies that next-generation evaluation systems can serve developmental purposes.[4] Rrecent research finds that professional development exhibiting certain characteristics can improve student achievement (Darling-Hammond et al, 2017; Donaldson, 2021). Next-generation evaluation systems, and NEE specifically, include several characteristics resembling effective professional development.

A recent literature review of 35 studies concerning teacher professional development finds that effective professional development exhibits one or more of the following: 1) teacher engagement in *active learning*, 2) *support for collaboration*, 3) models effective, *research-based practices*, 4) includes *coaching and expert support*, 5) offers teachers *feedback* and space for *personal reflection*, and 6) is *sustained* over time (Darling-Hammond et al., 2017). Conceptually, next-generation evaluation incorporates several of these characteristics. For example, teachers engage in *active learning* by identifying performance goals during structured observation conferences. Structured conferences also provide time for evaluator *feedback and reflection*; and, next-generation evaluation draws on aspects of *coaching and expert support* and *sustained* learning opportunities through recurring observations.

Next-generation evaluation's connection with coaching in particular may represent a potent professional development opportunity resulting in higher student achievement scores. Like evaluation's repeated observations and structured conferences, coaching programs provide

[4] There is an ongoing conceptual debate pitting "evaluation for development" against "supervision." Some argue that these are distinct (Firestone, 2014; Glickman et al., 2018; Mette et al., 2017), while a growing body of work argues that the two concepts share more in common than not (Donaldson, 2021; Papay, 2012; Woulfin & Rigby, 2017). We adopt the latter view.

teachers with ongoing, content-specific feedback to improve their effectiveness (Kraft, Blazar, & Hogan, 2018). A recent meta-analysis of the causal evidence corroborates this hypothesis: coaching's average impact on student achievement scores is as large as any known school improvement intervention (Kraft, Blazar & Hogan, 2018).

Study Context: Comparing the Performance-Based Teacher Evaluation system and Network for Educator Effectiveness

From the early 2000s through 2012-13, all Missouri districts implemented the Performance-Based Teacher Evaluation system (PBTE; for details see Missouri Department of Elementary and Secondary Education, 1999). In the early 2010s, researchers at the University of Missouri's College of Education developed NEE, a next-generation teacher evaluation system. NEE focused its development on rural users, seeking input from PreK-12 rural practitioners. A cohort of six rural districts volunteered to pilot NEE during 2011-12. The following school year, a second cohort of 26 more rural districts voluntarily joined. As a university-based nonprofit, NEE charges districts a fee of approximately $3 per student to recover operational costs. PBTE and NEE prioritized evaluation for teacher development, with the ultimate goal of improving student outcomes, and neither emphasized evaluation for accountability. Throughout 2011-12 and 2012-13, all non-NEE districts continued using PBTE.

In 2011-12 and 2012-13, the Missouri state education agency held meetings with dozens of districts and charter agencies to discuss proposed revisions to the state evaluation system that would be implemented after 2012-13 (Katnik, 2014). These reforms included using a revised Missouri-standards-based teacher performance rubric for formal evaluations, though the state did not mandate the use of a specific rubric. While evaluators (i.e., school administrators) were to evaluate teachers using an appropriate rubric, evaluators did not have to use the rubric for classroom observations. Indeed, Missouri's reforms did not require classroom observations for evaluative purposes but required at least "evidence of teacher performance" (Katnick, 2014). The reforms also called for evaluator training and certification but did not offer specific expectations. While not referenced by Missouri reforms explicitly, the reforms created d) a need for new teacher performance data management systems and e) analyses of those data for teacher human capital decision-making. Finally, state and local education agencies prioritized evaluation for development over accountability, implying a need to f) link evaluation and teacher professional development systems. Faced with these impending reforms, local education agencies had to develop and maintain (a) - (f) and any other teacher evaluation revisions adopted by local agencies or outsource these tasks to an external agency. Those districts choosing NEE decided to outsource these tasks to NEE for the nominal fee of $3 per student.

Some district and charter agency leaders that did not join NEE during the study period encouraged a few of their evaluators and teachers to test some of the state-agency-proposed reforms. In 2012-13, a total of 566 teachers across the state participated in this informal pilot and no district or school implemented the pilot systematically. Consistent with current state and NEE leadership, we assume that the 2012-13 state pilot does not represent a threatening form of treatment diffusion or contamination among PBTE schools.

We contrast PBTE and NEE using Liu and colleagues' framework (2019), defining evaluation systems according to a) rating specifications, b) sampling, and c) scoring procedures. Rating specifications describe observation protocols (i.e., rubrics) and sampling procedures include the number of performance indicators evaluators score for each observation, observation length, and observation frequency. Scoring procedures describe how evaluators generate scores. We also describe a) evaluator preparation and certification, b) observation conferences, and c) purposeful links between evaluation and professional development systems as prior work suggests that evaluation's success may depend on these elements (Donaldson, 2021). Table 1 summarizes the comparisons.

Observation Protocols

PBTE included an observation protocol describing six broad teacher performance standards (e.g., use of assessment for student learning, teacher content knowledge) and 20 finer-grain performance criteria embedded across the standards. Districts had the option to use a version of the protocol that described each performance criteria in terms of four different performance levels (Exceeds, Meets, Progressing, Does Not Meet), resulting in 80 different level-specific finer-grain descriptions. PBTE also allowed districts to adopt a three-point rating scale (Meets Expectations, Progressing Toward Meeting Expectations, Does Not Meet), but did not provide a protocol describing level-specific performance criteria. Finally, districts could develop their own protocols if they assessed teacher performance regarding the six performance standards and 20 performance criteria.[5]

NEE includes an observation protocol describing *research-based instructional practice* (Marshall, 2013), similar to Danielson's ubiquitous Framework for Teaching aligned with Missouri's teacher performance standards. NEE's protocol includes many performance criteria, each described in level-specific terms; in this way, NEE's protocol resembles the PBTE four-point protocol. However, all NEE districts use its protocol, while PBTE districts might not have adopted the four-point protocol with level-specific descriptions of performance criteria. NEE's protocol uses a five-point scale.

Number of Performance Criteria to Score

[5] The Missouri state education agency did not collect information about which scale or protocols districts used in the PBTE era.

PBTE evaluators judged teachers on one, two, or six performance criteria per observation. However, the rationale for these numbers and when they were applied is unclear.

NEE evaluators observe a teacher with respect to three to five performance criteria per observation. NEE encourages evaluators to choose several criteria that they can manage during observations while providing teachers with useful post-observation feedback for improvement. Furthermore, NEE exhibits the *active engagement* component of effective professional development as teachers *collaboratively* work with administration to select their yearly goals which, in turn, influence the criteria upon which they are evaluated. NEE teachers are also expected to *actively engage* with their post-observation feedback and *collaborate* with colleagues, coaches, or administrators to improve their performance.

Observation Length, Frequency, and Conferences

PBTE policy documents recommended that teachers in their first three years on the job receive one scheduled (i.e., announced) and two unscheduled (i.e., unannounced) observations per year. Pre-tenure teachers beyond their third year were recommended to receive one scheduled and one unscheduled observation per year, and tenured teachers were to receive one observation during their formal evaluation year only. The PBTE did not specify how long an observation should last.

NEE characterizes its observations as “short mini-observations” and recommends that all teachers receive six to ten mini-observations per year. In other words, NEE treats its observations as a *sustained* learning opportunities throughout the academic year that maintain *active engagement* on the behalf of teachers. Furthermore, PBTE and NEE expected evaluators to hold a conference after each observation during which evaluators shared performance feedback and developed teacher improvement plans, providing opportunities for *feedback* and *reflection*, a characteristic of teacher coaching (Kraft, Blazar & Hogan, 2018) and effective professional development (Darling-Hammond et al., 2017).

Observer Preparation and Certification

PBTE policy documents did not describe systematic evaluator preparation programs, expectations or describe evaluator credentialing or certification.

NEE evaluators receive annual and ongoing NEE-provided 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, the latter of which improves teacher value-added to student achievement scores (Cravens & Hunter, 2021). Following training, prospective evaluators must pass a certification exam each summer to receive certification to conduct formal observations.

Expected Changes in Student Achievement

Switching from PBTE to NEE is expected to increase student achievement scores for several reasons. All NEE districts adopted a standards- and *research-based observation protocol describing instructional practices*, while PBTE districts might have done so. Although the extent to which PBTE teachers were actively engaged in the selection of their professional learning goals is unclear, NEE teachers *actively engage* in this selection process and in their improvement via post-observation conferences. Moreover, NEE evaluators are trained to *collaborate* with teachers directly and support teacher *collaboration* with other personnel to improve instruction. NEE teachers are also assigned more frequent observations and post-observation feedback conferences, providing NEE teachers *sustained* opportunities to receive performance-enhancing *feedback* and *reflect* upon it. Ultimately, the NEE observation process resembles *coaching*, one of school improvement's most potent interventions aiming to raise student achievement scores (Kraft et al., 2018). Additionally, NEE evaluators are certified annually and receive ongoing training, which represents characteristics of effective professional development for evaluators, a key lever for effective teacher evaluation (Steinberg & Sartain, 2015).

As school districts implement NEE and PD-adjacent features, it is important to consider potentially moderated effects across teachers and school characteristics, as implementation of teacher evaluation varies by setting (Donaldson & Woulfin, 2018; Marsh et al, 2017). We consider and examine this potential heterogeneity in school-level average prior year achievement, school-level average teacher experience, school-level concentration of nonwhite students, and school-level concentration of FRPL students. Ultimately, we hypothesize that less-advantaged school settings will benefit more from NEE's implementation and strong focus on development. Not all teachers and students have the same growth potential, so it follows that those with more room to grow will benefit more from NEE's developmental features.

Finally, we explore effects by NEE cohort and over time within one cohort. As described in further detail below, we have data for NEE's first two cohorts. Although we prefer to "pool" the cohorts together to increase power and estimate NEE's effects one year after each cohort's implementation, we also examine whether one cohort or the other drives NEE's effects. We also leverage the two years of data for NEE's first cohort to explore if NEE's effects change over time.

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, but not student-to-teacher links. 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.

Methods

Our primary estimation goal is identifying the causal impact of introducing NEE on math and reading achievement scores one year after implementation. Although evaluation's effects might take more than one year to materialize, empirical evidence suggests otherwise (Steinberg & Sartain, 2015; Taylor & Tyler, 2012). Ideally, the research design would compare NEE districts' post-implementation achievement scores to the scores NEE students would have generated in the absence of treatment. As the latter are unobservable, causal inference depends on identifying comparison scores approximating the NEE counterfactual. We apply a difference-in-difference (DID) strategy and compare deviations from prior achievement score trends among students in NEE districts to corresponding deviations for students in matched PBTE districts. To identify a valid comparison group, we identify matching PBTE districts whose pre-intervention achievement trends resembled NEE districts' pre-intervention trends. Post-intervention deviations in achievement trends between NEE and matched comparison districts with similar pre-intervention trends are attributed to NEE's introduction.

We use coarsened exact matching (CEM) to match districts, coarsening the data into *strata* per Sturge's Rule. CEM then identifies strata with NEE districts and identifies within-strata PBTE matches. The CEM uses historical achievement scores at the district level, urbanicity, and district historical PPE as matching variables; districts are the units of analysis in the matching procedure as selecting into a Missouri evaluation system is a district decision. At a minimum, CEM should match on historical achievement as DID internal validity largely rests on parallel historical achievement trends between NEE and comparison districts. We also match on urbanicity and PPE because NEE is fee-based and designed for rural districts specifically.

Matching occurs by cohort because NEE's implementation was staggered over time. The pool of potential matches for Cohort 1 includes all districts that continuously used PBTE through 2011-12, the year NEE launched in Cohort 1. Districts that implemented NEE in 2012-13 were also in Cohort 1's pool of potential matches. The CEM procedure matches on four historical district-level average student achievement score variables: historical scores one, two, three, and four years before 2011-12 (i.e., 2007-08 - 2010-11). The procedure also matches on four historical district-level PPE variables and 2011-12 urbanicity. Cohort 2's matching procedure is analogous to Cohort 1's except that the pool of potential matches includes all districts that continuously used PBTE through 2012-13. Then, 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); centered-years in the stacked data ranged from -4 to 0.

Following Gormley and Matsa (2011), we apply a generalized DID model to stacked data using Equation 1:

$$y_{isdtc} = \delta NEE_{dt} + \beta_1 y_{isd(t-1)} + \beta_2 PPE_{d(t-1)} + \beta_3 Rural_{dt} + \Delta_{dc} + \Phi_{tc} + e_{isdtc} \quad (1).$$

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. The independent variable of interest, NEE_{dt}, is an indicator equaling one for NEE districts after NEE's launch. Equation 1 applies district-cohort FE and year-cohort FE, effectively comparing deviations in achievement trends within each stack (Gormley & Matsa, 2011). Equation 1 also includes prior-year student achievement, prior-year district PPE, and urbanicity. By controlling for prior-year achievement scores δ plausibly represents the change in achievement scores NEE students experienced due to one year of NEE implementation in their district. Our preferred specification includes standard errors that are district-student-cohort multiway clustered. While the use of a matched comparison group bolsters internal validity, Equation 1's estimates may only apply to districts in NEE's first two cohorts (i.e., average treatment effect on the treated, ATT).

Sensitivity Tests

Our sensitivity tests begin by re-applying Equation 1 using a larger set of control variables. The larger set includes 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 years of experience; and school- and district-level average student prior-year achievement scores. To the extent DID identification assumptions are met, controls are unnecessary; however, the use of control variables is conventional. We find that NEE's ATT is insensitive to the use of these expanded controls.

Sensitivity tests also estimate versions of Equation 1 using a) the canonical district FE and year FE, b) district FE, year FE, and expanded controls, and c) district-cohort FE, year-cohort FE, and cohort-specific expanded controls; tests a) – c) generate similar effects.

Finally, as prior work in urban settings examines moderated effects by teacher and school characteristics, we estimate similar effects by interacting a continuous variable measuring the school-level average student prior-year achievement score, school-level average teacher years of experience, school-level concentration of FRPL students, or school-level concentration of nonwhite students with treatment.

Internal Validity

Unbiased estimation of δ is threatened if unobserved factors systematically influenced student achievement a) at the same time NEE launched in cohort 1 or 2 and b) these influences differed by evaluation system (i.e., NEE, PBTE). Indirect evidence from institutional knowledge, parallel trend tests, and placebo tests can mitigate these violations' plausibility, but analysts cannot test for direct violations of a) or b) directly. Although not required to meet DID identification assumptions, balance tests reveal the extent to which the DID quasi-experimental design 'randomized' units to treatment or control status. No evidence from any of these tests threatens the identification of δ.

Parallel Trends Test and Event Study Analysis. Event study analysis is used to explore pre-intervention parallel trends and estimate treatment effects nonparametrically. The event study analysis compares pre- and post-intervention student achievement in NEE and matched PBTE districts by each year preceding NEE's launch and the year of its launch in each cohort. Equation 2 describes the event study model:

$$y_{isdtc} = \delta_{-4}NEE_{dt} + \delta_{-3}NEE_{dt} + \delta_{-2}NEE_{dt} + \delta_0 NEE_{dt} + \beta_1 y_{isd(t-1)} + \beta_2 PPE_{d(t-1)} + \beta_3 Rural_{dt} + \Delta_{dc} + \Phi_{tc} + e_{isdtc} \quad (2)$$

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 represent 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 PBTE districts are relatively parallel over time, meeting a DID identification assumption, then δ_j will be statistically insignificant when j < 0. Additionally, δ_0 represents the ATT, corresponding with Equation 1's δ. Other terms refer to the same quantities as Equation 1.

Placebo Tests and Institutional Knowledge. Estimates δ or δ_0 may capture spurious effects of interventions implemented in the same year as NEE's respective launches in Cohorts 1 or 2. According to several sources with intimate knowledge of NEE and PBTE in the early 2010s, neither NEE cohort nor PBTE districts systematically implemented alternative confounding treatments in 2011-12 or 2012-13. NEE's founders, who remain its current leaders, worked closely with Cohort 1 and 2 district leaders. Indeed, NEE closely monitored Cohort 1 and 2 district activities to learn about NEE's implementation. Based on many meetings between NEE leaders and Cohort 1 and 2 district leaders, NEE's founders have no knowledge of any systematically implemented or plausibly confounding non-NEE interventions. Additionally, the then-Director and current Assistant Commissioner of DESE was intimately involved with Missouri districts' transitions from PBTE to a next-generation system in the early 2010s. The Assistant Commissioner also reports no knowledge of factors that systematically influenced student achievement in PBTE during 2011-12 or 2012-13. Moreover, Katnik (2014) details DESE's small-scale piloting of next-generation evaluation in the early 2010s. As discussed in the Study Context section, some district leaders encouraged a few evaluators and teachers to test some aspects of DESE's pilot. Only a total of 566 teachers across the state participated in this informal pilot; no district or school implemented the pilot systematically.

However, the ATT may be biased if interventions in the years preceding NEE's launch affected student achievement. 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 ≥ 2010-11;

centered-year ≥ -1). The remaining placebo tests similarly recode NEE_{dt} for the remaining false years of treatment.

Balance Tests. Baseline balance tests check the extent to which NEE students, schools, and districts are statistically indistinguishable from the comparison group regarding observable characteristics. Although DID identification assumptions do not require such balance as the research design absorbs between-district-cohort and between-year-cohort differences, balance in the observables may further the plausibility of causality. Online Appendix A describes the baseline balance methods in detail.

We also check the balance of characteristics measured during the year of NEE's launch. Measured characteristics include each variable from the expanded list of control variables discussed in the Sensitivity Test section. NEE was not designed to alter the composition of districts regarding student or teacher gender or race, student FRPL, or teacher education level or years of experience. Evidence of post-intervention imbalance may suggest that NEE and PBTE districts systematically implemented confounding interventions during NEE's launch.

Effects Over First Two Years: Cohort 1

Although the study's primary purpose is identifying the ATT after one year of implementation, Cohort 1's data allow for the identification of NEE on achievement scores one and two years after introduction. To estimate these dynamic effects, we retain Cohort 1 and its matched comparison group only. 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 records came from 2012-13 or not.

Falsification Test: Teacher Mobility Analysis

We use falsification tests to support our causal interpretation of NEE's effect on student achievement scores. NEE's focus on developmental evaluation, rather than high-stakes evaluation, means that we should not detect any effects of introducing NEE on teacher mobility outcomes. We examine the ATT of introducing NEE on teacher mobility using Equation 3:

$$m_{isdtc} = \delta NEE_{dt} + \beta_2 PPE_{d(t-1)} + \beta_3 Rural_{dt} + \Delta_{dc} + \Phi_{tc} + e_{isdtc} \quad (3),$$

where m_{isdtc} represents one of two teacher mobility indicators. First, we operationalize district-switchers such that m_{isdtc} is 1 in year *t* if a teacher works in a different Missouri public school district in year *t+1*. Second, we operationalize Missouri public school systems exits such that m_{isdtc} is 1 in year *t* if a teacher is no longer employed by a Missouri public school district in year *t+1*. All other terms in Equation 3 are defined identically to Equation 1.

Findings

Pre-Matched Descriptive Statistics

Aside from differences in student race, urbanicity, and PPE, NEE districts resemble the sample of all (i.e., matched and unmatched) PBTE districts (see Table 2). While 22 percent of PBTE students are nonwhite, just 11 percent of NEE students are nonwhite, which is explained by the urbanicity of NEE and PBTE districts. Indeed, this is the starkest difference between NEE and PBTE districts: all NEE districts are rural (i.e., "rural" or "town" per NCES), while 84 percent of PBTE districts are rural. Finally, the average NEE district spends about $1,500 less per pupil, mitigating concerns that districts choosing to pay NEE's nominal fee are wealthier.

Matching Results

As the validity of our strategy does not depend on post-matching covariate baseline balance at the district level for the reasons above, 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 PBTE districts. Cohort 2 matching used 287 coarsened strata, matched using 16 strata, and matched 19 of 26 NEE districts to 127 PBTE districts. The mean differences between matched NEE and PBTE 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 with 120 PBTE districts while 24 Cohort 2 districts matched with 197 PBTE 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 including rural districts only (for further details, see Online Appendix B).

District-Level Prior-Year Student Achievement Trends

There is some evidence that pre-intervention achievement trends in districts that remained in PBTE 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 average students' achievement scores in NEE, PBTE, and matched PBTE districts. The top-left panel suggests that PBTE and Cohort 1's pre-intervention math score trends are not parallel. While PBTE 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 largely parallels the PBTE trend. The matching procedure resulted in matched prior-year math score trends that largely parallel NEE trends in each cohort. Moreover, Cohort 1's matched PBTE district pre-intervention trends are not only parallel but near-equivalent. The bottom-left panel shows that NEE, all PBTE, and matched PBTE pre-intervention trends are largely parallel, although NEE district reading scores deviate from the trend four years prior to 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 DID design rests on parallelism in *student*-level pre-intervention trends as students are the unit of analysis in the DID. We examine the parallelism of pre-intervention student-level achievement trends in NEE and matched PBTE districts via event study analysis.

Parallel Trends Test and Event Study Results

Event study results show that pre-intervention achievement score trends are consistent with the parallel trends assumption and suggest that NEE improved achievement scores slightly, but not by a statistically significant amount (Figure 2). Pre-intervention differences in achievement across NEE and matched PBTE districts are individually and collectively[6] statistically indistinguishable from the score difference in the year before NEE's launch (i.e., pre-intervention confidence intervals overlap with zero); thus, pre-intervention trends are parallel.

[6] Although not a requirement of event study tests, we estimate the joint significance of pre-intervention estimates by subject, a much more rigorous test than is conventional. The joint significance of math (reading) pre-intervention estimates is $p \sim 0.11$ ($p \sim 0.42$), furthering our confidence in the parallel trends assumption.

The bottom coefficient in each panel of Figure 2 shows that NEE's launch increased math and reading achievement scores by 0.01 standard deviations (SD) relative to the year before the intervention, though the change is not statistically significant.

Generalized DID Results

NEE's ATT on math and reading scores are consistent with the event study results, insensitive to model specification, and not moderated by cohort. Column I of Table 3 shows that the generalized DID ATTs on math and reading scores are 0.01 SD but not statistically significant, resembling the event study estimates. Equation 1's ATTs are not sensitive to use of the expanded set of controls (column II), cohort-specific controls (column III), replacement of district-cohort FE and year-cohort FE with district FE and year FE (column IV), nor the use of the expanded controls with district FE and year FE (column V). Indeed, the ATT is consistently 0.01 SD in each subject. Furthermore, Column VI results, which moderate the ATT by cohort, find no evidence of moderation across cohorts and cohort-specific estimates also resemble nonmoderated effects.

Despite statistical insignificance, we are not concerned with low power and instead conclude that introducing NEE does not impact achievement scores on average. We reach this conclusion based on Jacob et al (2019) and Kraft's (2020) interpretation of null findings. All point estimates in our main findings (Table 3) are small, between 0.01 and 0.02 SD. Confidence intervals (CIs) are also precisely estimated and as low as 0.00 to 0.03 SD. Moreover, a small effect size is less than |0.05| SD according to Kraft (2020). Using this framework, all point estimates are considered small as are nearly all of the CI bounds.

Effects Over First Two Years: Cohort 1

Results in Table 4, which capture Cohort 1 effects one and two years after NEE's introduction, are strikingly similar to the main, sensitivity, and cohort-moderated effects in Table 3. NEE's Cohort 1 ATT on math scores one year after implementation is 0.01 SD, as is its ATT two years after introduction, yet neither is statistically significant (Panel A, Table 4). Moreover, the point estimates and confidence intervals are identical, ruling out differential effects on math scores over time. Cohort 1 reading effects over time exhibit a similar pattern (Panel B). NEE's ATTs are 0.01 SD, and while the confidence intervals are not identical, they overlap substantially. Ultimately, the evidence suggests that NEE's ATTs may not change over time.

Internal Validity

Placebo and balance tests affirm the research design's internal validity. Placebo tests produce no evidence of false treatment 'effects' on scores in either subject during any pre-intervention year (Table 5). Most 'effects' in Table 5 are less than 0.01 SD or negative, confirming that we do not observe effects when we expected none. Moreover, Column II (III) shows that two (three) years before NEE's launch, math (reading) achievement scores rose by a statistically insignificant 0.02 (0.01) SD, which is at least as large as the post-NEE changes in student achievement. These false ATTs reinforce the conclusion that NEE had no discernible effect on achievement scores. In NEE's absence throughout the pre-intervention years, achievement scores changed just as much as they did after NEE's launch.

Baseline balance tests suggest that the quasi-experiment effectively 'randomized' NEE students, schools, and districts to treatment. Results in Table 6 show that prior-year achievement scores at the student-, school-, and district-level, and district prior-year PPE, balanced across NEE and matched PBTE districts, underscoring the comparability of these two groups. Indeed, differences in prior-year achievement scores at each level, perhaps the most important baseline characteristic to 'randomize,' are virtually zero.

'Effects' on characteristics other than prior-year achievement and PPE also suggest the absence of confounding treatments during the year of NEE's launch. The remaining balance tests find that student and teacher race and gender, student FRPL, and average teacher education level and years of experience were unaffected during the year of NEE's launch (Table 6).

Falsification Test: Teacher Mobility Analysis

We do not find evidence that NEE affected either measure of teacher mobility, district switches nor exits from the Missouri public school system. This supports our causal interpretation of effects on student achievement as NEE is designed to be developmental; it is not implemented for the purposes of personnel decision making (e.g., teacher dismissal). Table 7 reports the ATT on district switches using the math sample only. Using our preferred model, we estimate a statistically insignificant ATT of -0.03 percentage points (Panel A1, Column I). The ATT on switching districts is not sensitive to additional teacher-level controls (column II), cohort controls (column III), use of district FE and year FE in lieu of district-cohort FE and year-cohort FE (column IV), nor expanded teacher-level controls with district FE and year FE (column V). Furthermore, while effects on district mobility switches from negative to positive between cohorts, effect sizes are still insignificant and therefore there is no evidence of heterogeneity of ATTs across cohorts (column VI).

Similarly, we do not detect effects teachers exiting the Missouri public school system. Across all pooled models (Panel B1, Columns I-V), we find an ATT of zero or near-zero percentage points. Again, there is no evidence of heterogeneity across cohorts (Panel B2, Column VI). All mobility analysis results are insensitive to the use of the reading sample (Online

Appendix Table C1). Overall, no evidence suggests that ATTs on student math and reading achievement are driven by teacher turnover, consistent with NEE's developmental purpose.

Moderation Analyses

Although there are no discernible average effects, NEE increases math and reading scores in disadvantaged schools, sometimes substantially. Figures 3 and 4 graph NEE's total effects (i.e., $\delta_{m1}+\delta_{m2}$) on math and reading scores, respectively. The abscissae of each panel ranges from each moderator's 5th to 95th percentile.[7] With the exception of school-level FRPL concentration (top-right panels, Figures 3, 4), the other school characteristics moderate NEE's impact in at least one subject area. As FRPL is not a moderator, we do not discuss it further.

Schools with low prior-year math achievement scores benefit from NEE (top-left panel Figure 3). NEE's effect on math scores in schools with average student prior-year math achievement scores of -0.1 SD and below are significantly higher than effects in schools where prior-year scores are 0.15 SD and above. Further, NEE significantly increases math scores by 0.03 to 0.01 SD in schools where the average student's prior-year math score was at or below the statewide average score (i.e., ≤ 0). However, the data also show that NEE may negatively affect math scores in high-performing schools.

The bottom-left panel of Figure 3 reveals that NEE's impact on math scores rises with the concentration of nonwhite students in a school and improves achievement scores by 0.01 to 0.03 SD in schools where more than 5 percent of students are nonwhite. The effects in schools where 20 percent or more of students are nonwhite exceeds the effect in schools with no nonwhite students by a margin of 0.02 SD.

[7] The percentiles of moderators in the math sample differ from percentiles in the reading sample because the matched samples differ.

Average teacher experience moderates NEE's effects on math scores substantially, with the effect declining as the average teacher gains experience. NEE's largest detected impact on math scores occurs in schools with the least experienced average teacher (5 years; ATT ~ 0.12 SD) and rises just over 0.01 SD for a one-year *decline* in the years of experience held by a school's average teacher (Figure 3 bottom-right panel). NEE's impact remains significant and positive until the school-level average teacher's years of experience reaches about 13 years, the years of experience for the average teacher statewide, at which point the ATT becomes statistically insignificant.

The moderated effects on reading achievement scores resemble math effects as average student prior-year reading achievement moderates NEE's impact negatively (top-left panel, Figure 4). Schools with reading achievement scores below the statewide average benefit from NEE and effects are discernibly different between schools where average student prior-year scores differ by more than 0.25 SD (e.g., NEE's effect in schools where the average reading score is -0.10 SD is statistically higher than schools where the average score is 0.15SD). However, unlike the math results, there is no evidence that NEE may negatively affect reading scores in high-performing schools.

Again, NEE is most effective in schools where the typical teacher is less experienced, though not as effective in raising reading scores as raising math scores. In schools where the average teacher is below the state average (i.e., 13 years), NEE's effect on reading scores is positive, ranging to approximately 0.04 SD. Extrapolating the experience-moderator trend line to five years of average teacher experience, the minimum in the corresponding math-sample graph, suggests that NEE's impact on reading scores is 0.05 SD. Thus, NEE's impact on math scores in schools with less-experienced average teachers is substantially greater than its impact on reading scores in similar schools. Similarly, for each one-year *decline* in average teacher years of experience, NEE's effect rises by approximately 0.005 SD, less than half the increase in NEE's effects on math scores.

Similar to effects on math scores, the concentration of FRPL students in schools is not a moderator (top-right panel, Figure 4); however, unlike the math effects, neither is the nonwhite student moderator (bottom-left panel, Figure 4).

Conclusion

Experimental and quasi-experimental research from urban and national settings find mixed evidence concerning the introduction of next-generation teacher evaluation systems on student achievement scores (Bleiberg et al., 2021; Steinberg & Sartain, 2015; Taylor & Tyler, 2012). However, no rigorous research identifies the plausibly causal effects of next-generation evaluation in rural settings, although more than half of school districts in the United States are rural (National Center for Education Statistics, 2013), and evidence suggests that evaluations' effects may vary by urbanicity (Rodriguez et al., 2020). Moreover, education policymakers crafting teacher evaluation policies for rural settings may prioritize rurally-situated research over internally valid studies in non-rural settings (Nakajima, 2022). The current study addressed these gaps by applying a difference-in-differences (DID) framework to rural Missouri administrative data from 2007-08 through 2012-13, identifying the plausibly causal effects of the Network for Educator Effectiveness (NEE), a next-generation teacher evaluation system, on math and reading achievement scores.

As NEE is fee-based, we discuss its effects and effects-to-expenditure ratios, a novel contribution to the teacher evaluation literature. Ideally, we would prefer to describe NEE's net costs or cost-effectiveness, because expenditures do not capture all relevant costs. For example, suppose that policymakers could adopt NEE or another intervention shown to have similar effects on student achievement. Furthermore, suppose that both NEE and the other intervention cost districts $3 per student; however, the other intervention requires far more school administrator training than NEE, *ceteris paribus*. The effect-to-expenditure ratios for NEE and the other intervention will be similar, but NEE is more cost effective. We presume that policymakers prefer cost-effectiveness ratios over effect-to-expenditure ratios. We also presume that policymakers prefer effect-to-expenditure ratios over the discussion of effects only, as the former affords some formal sense of effects and cost.

We conclude that NEE did not affect student math or reading achievement, on average. The average treatment effects on the treated (ATTs) are robust to several sensitivity tests, do not vary by cohort, and do not change in the second year of Cohort 1's implementation. Importantly, effects in this time frame are plausible as prior work has shown statistically significant effects for similar interventions in urban settings after just one year of implementation (Steinberg & Sartain, 2015; Taylor & Tyler, 2012). If we interpret the precisely estimated null effects to mean that NEE has no effect on achievement scores, the effects-to-cost ratio is zero. To place the ratio of zero 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 0.000003SD in mathematics.

Despite the main null findings, we conclude that NEE's introduction increased student achievement in math and reading, sometimes substantially, in disadvantaged rural settings. First, school-level average student prior-year scores moderated ATTs in each subject. NEE increased math and reading scores in rural schools with prior-year achievement scores at or below the state average, with effects ranging from approximately 0.015SD to 0.03SD. Importantly, the largest effects were in the lowest-performing schools and are equivalent to approximately one month of learning.[8] The effects-to-expenditure ratios in these schools range from 0.0005SD to 0.01SD per dollar spent, substantial returns to dollars spent.

Rural schools with higher concentrations of nonwhite students also benefitted from NEE's introduction. Math, but not reading, scores increased in virtually all NEE schools with any nonwhite students that adopted NEE, and the effects rise with the concentration of nonwhite students, improving math scores by as much as 0.03SD or 0.01SD per dollar spent. The importance of this finding extends beyond money well-spent; prior research shows that White-Black, White-Hispanic, and White-Native American achievement gaps persists in rural schools (Johnson et al., 2020). Our finding suggests that NEE can shrink this gap.

Although achievement scores rose by as much as 0.03 SD in low-performing and high-minority rural schools, NEE's most substantial effects are in rural schools with less-experienced teachers. Rural schools where the average teacher's years of experience are below the state average (13 years) benefit from NEE, and the effects are strongest in schools with the least experienced average teacher. Indeed, NEE improves math achievement scores up to 0.12 SD, or four months of learning, in schools with the least experienced average teacher, similar to the meta-analytic causal effects of instructional coaching on student achievement (Kraft et al., 2018). The effects-to-expenditure ratios in rural schools where the average teacher is below the state average range from approximately 0.0005SD to 0.04SD per dollar spent, which is staggering.

---

[8] 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.

Ultimately, the moderation of NEE's effects is consistent with causal inferences. At its core, NEE aims to improve student outcomes by developing teaching, and the current study found effects in schools with the most potential for improvement. Research consistently shows that teachers with less experience and those teaching lower-achieving and nonwhite students are typically less effective (Clotfelter et al., 2005, 2006; Goldhaber et al., 2015; Ladd & Sorensen, 2017; Papay & Kraft, 2013). NEE's effects are the largest in these settings.

Although the current study used moderators similar to those in Steinberg and Sartain's study of Chicago teacher evaluation (2015), their results differ from ours substantively. School-level prior-year achievement positively moderated the effects of Chicago's next-generation evaluation system but negatively moderated NEE's effects. NEE's effects also interacted with moderators that did not moderate Chicago's effects and vice versa. Lower-poverty schools benefited more from Chicago's system than higher-poverty schools, but school-level poverty did not moderate NEE's effects. However, school-level average teacher years of experience negatively moderated NEE's effects in both subjects while the concentration of nonwhite students in a school positively moderated the ATTs on math scores; however, these characteristics did not moderate Chicago's impact. It is unclear why the developmentally focused Chicago and rural Missouri teacher evaluation systems generate such different moderated effects. At face value, urbanicity may be the explanation, but research should test this conjecture.

Limitations

This study may be limited in several ways. First, the estimates may not capture the change in student achievement a typical PBTE district would have observed if it switched from PBTE to NEE (i.e., we assume the research design generated ATTs).

Second, the ATTs may not generalize to other settings; indeed, the results may be restricted to rural settings. Even findings generated by urban-situated studies have not transferred across cities; Cincinnati's evaluation system produced effects on math scores only (Taylor & Tyler, 2012) while Chicago's affected reading scores only (Steinberg & Sartain, 2015). Furthermore, the effects of an evaluation system may also depend on design (e.g., observation frequency, observer training) and purpose (Donaldson, 2021).

Third, NEE's ATTs may change over longer time periods. Although analyses of Cohort 1's ATT after one and two years of implementation did not imply a growth trajectory, longer panels could explore if NEE's effects increase as districts gain experience with the system.

Fourth, we only examine student achievement outcomes, but NEE may affect other student or educator outcomes. Indeed, we assume that NEE's users, particularly those in relatively advantaged rural schools, believe it affects important unexamined outcomes positively; otherwise, we cannot fathom why education agencies overseeing these schools 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 previously, we report effects-to-expenditure ratios, falling short of the ideal cost-effectiveness ratios.

Implications

Although the collective evidence concerning the introduction of teacher evaluation systems leans towards no detectable effects, on average, it implies that there are settings in which evaluation improves achievement. In this regard, the evidence from Cincinnati (Taylor & Tyler, 2012) and rural Missouri is consistent: evaluation improves achievement in settings with substantial improvement potential. However, the Chicago system produced Matthew effects, whereby advantaged schools benefitted the most (Steinberg & Sartain, 2015). We interpret the collective evidence to mean that introducing a next-generation evaluation system may benefit disadvantaged schools. However, future work should test this interpretation, especially when considering the evidence from Chicago. Although some excellent work has examined the conditions under which evaluation works (e.g., Donaldson & Woulfin, 2018; Marsh et al., 2017), we argue that scientists and practitioners alike need more information in this arena.

Our work also affords targeted policy implications, which we offer while urging the caution befitting implications stemming from a single study. To the extent our results generalize to other settings, the evidence suggests it may be advantageous for rural districts without a next-generation teacher evaluation system to adopt one, especially if the district includes a sizeable number of disadvantaged schools. However, we strongly recommend that an adopted system mimic NEE by adopting: a rubric-based protocol for observations; frequent, structured, and short observations; and extensive training for school leaders.

To this end, we emphasize that NEE is a university-based non-profit developed by scholars with ongoing input from end-users; we also speculate that it would be difficult for rural education agencies to replicate NEE's services. While we assume that many education agencies have developed and refined teacher evaluation with some scholarly input, we believe that meaningful and engaging researcher-practitioner partnerships like NEE can yield effective teacher evaluation practices in disadvantaged rural schools. NEE's story also implies that rural education agencies trust university-based teacher evaluation systems and value these systems above nominal fees; otherwise, we presume these agencies would choose self-designed evaluation systems.

We also speculate that rural users might value such researcher-practitioner partnerships due to capacity constraints. While large districts might employ offices or individuals providing NEE-like services, it would be difficult for rural (i.e., substantially smaller) education agencies to do the same. Instead, the rural education agent responsible for teacher evaluation might also manage several other schooling operations, crowding out the time rural agents can devote to evaluation for development. A university-based partnership can expand rural district capacity substantially via measurement development, evaluation data management and analysis, direct technical assistance and support, and the review and incorporation of research-based practices in evaluation systems. Indeed, the expertise and time NEE offers rural districts might explain its positive effects in disadvantaged rural schools.

Finally, NEE's effects suggest that rural education agencies can use it to improve student achievement in disadvantaged schools and that NEE's fee is money well spent. States policymakers might incentivize disadvantaged rural schools to implement NEE-like systems by assigning state-provided funds in these specific schools for next-generation systems like NEE. Effect-to-expenditure ratios imply that it makes little sense for state policy to do the same for advantaged rural schools; however, such policy might lead advantaged rural schools to leave NEE, reducing NEE's income. While these losses would affect the scope of NEE's work, it is unlikely they would lead NEE to lay off critical staff or discontinue essential services as NEE staff are full-time university faculty and staff. Indeed, this underscores another benefit of a fee-based, non-profit researcher-practitioner partnership situated within a university.

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/

Archer, J., Cantrell, S., Holtzman, S., Joe, J., Tocci, C., & Wood, J. (2016). *Better feedback for better teaching: A practical guide to improving classroom observations* (1st ed.). Jossey-Bass.

Bleiberg, J., Brunner, E., Harbatkin, E., Kraft, M. A., & Springer, M. G. (2021). *The Effect of Teacher Evaluation on Achievement and Attainment: Evidence from Statewide Reforms* (Working Paper No. 21–496; EdWorkingPaper). Annenberg Institute at Brown University. https://www.edworkingpapers.com/ai21-496

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.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. (2005). Who teaches whom? Race and the distribution of novice teachers. *Economics of Education Review*, *24*(4), 377–392. https://doi.org/10.1016/j.econedurev.2004.06.008

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.

Cohen, J., Loeb, S., Miller, L. C., & Wyckoff, J. H. (2020). Policy Implementation, Principal Agency, and Strategic Action: Improving Teaching Effectiveness in New York City Middle Schools. *Educational Evaluation and Policy Analysis*, *42*(1), 134–160. https://doi.org/10.3102/0162373719893338

Cravens, X. C., & Hunter, S. B. (2021). Assessing the impact of collaborative inquiry on teacher performance and effectiveness. *School Effectiveness and School Improvement*, *Online*, 1–43. https://doi.org/10.1080/09243453.2021.1923532

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

Firestone, W. A. (2014). Teacher Evaluation Policy and Conflicting Theories of Motivation. *Educational Researcher*, *43*(2), 100–107. https://doi.org/10.3102/0013189X14521864

Glickman, C., Gordon, S., & Ross-Gordon, J. (2018). *Supervision and Instructional Leadership* (10th ed.). Pearson.

Goldhaber, D., Lavery, L., & Theobald, R. (2015). Uneven Playing Field? Assessing the Teacher Quality Gap Between Advantaged and Disadvantaged Students. *Educational Researcher*, *44*(5), 293–307. https://doi.org/10.3102/0013189X15592622

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.

Hattie, J., & Timperley, H. (2016). The Power of Feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hunter, S. B., & Rodriguez, L. A. (2021). Examining the demands of teacher evaluation: Time use, strain and turnover among Tennessee school administrators. *Journal of Educational Administration*, *59*(6), 739–758. https://doi.org/10.1108/JEA-07-2020-0165

Hunter, S. B., & Springer, M. G. (2021). Performance Feedback, Human Capital, and Teacher Performance: A Mixed-Methods Analysis. *Educational Evaluation and Policy Analysis*, *Online*. https://journals.sagepub.com/eprint/KH35P64VZKPK33BZRNBT/full

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), 36.

Katnik, P. J. (2014). *A Study of Missouri's Educator Evaluation System and its Efforts to Increase Teacher and Leader Effectiveness* [Dissertation, University of Missouri, St. Louis]. https://irl.umsl.edu/cgi/viewcontent.cgi?referer=&httpsredir=1&article=1254&context=dissertation

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., Blazar, D., & Hogan, D. (2018). The Effect of Teacher Coaching on Instruction and Achievement: A Meta-Analysis of the Causal Evidence. *Review of Educational Research*, *88*(4), 547–588. https://doi.org/10.3102/0034654318759268

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.

Ladd, H. F., & Sorensen, L. C. (2017). Returns to Teacher Experience: Student Achievement and Motivation in Middle School. *Education Finance and Policy*, *12*(2), 241–279. https://doi.org/10.1162/EDFP_a_00194

Liebowitz, D. D., Porter, L., & Bragg, Dylan. (2019). *THE EFFECTS OF HIGHER-STAKES TEACHER EVALUATION ON OFFICE DISCIPLINARY REFERRALS* (Working Paper No. 19–159; EdWorkingPaper). Annenberg Institute at Brown University. http://www.edworkingpapers.com/ai19-159

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

Liu, S., Bell, C. A., Jones, N. D., & McCaffrey, D. F. (2019). Classroom observation systems in context: A case for the validation of observation systems. *Educational Assessment, Evaluation and Accountability*, *31*(1), 61–95. https://doi.org/10.1007/s11092-018-09291-3

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

Mette, I. M., Anderson, J., Nieuwenhuizen, L., Range, B. G., Hvidston, D. J., & Doty, J. (2017). The Wicked Problem of the Intersection between Supervision and Evaluation. *International Electronic Journal of Elementary Education*, *9*(3), 709–724.

Missouri Department of Elementary and Secondary Education. (1999). *Guidelines for Performance-Based Teacher Evaluation*. Missouri Department of Elementary and

Secondary Education. https://web.archive.org/web/20061005043323/http:/dese.mo.gov/divteachqual/leadership/profdev/PBTE.pdf

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 Center for Education Statistics. (2013). *The Status of Rural Education* (Spotlight Chapter 3; p. 7). NCES.

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

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

Papay, J. P., & Kraft, M. A. (2013). Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement. *Journal of Public Economics*, *130*, 105–119. https://doi.org/10.1016/j.jpubeco.2015.02.008

Phipps, A. R., & 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

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* (No. 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

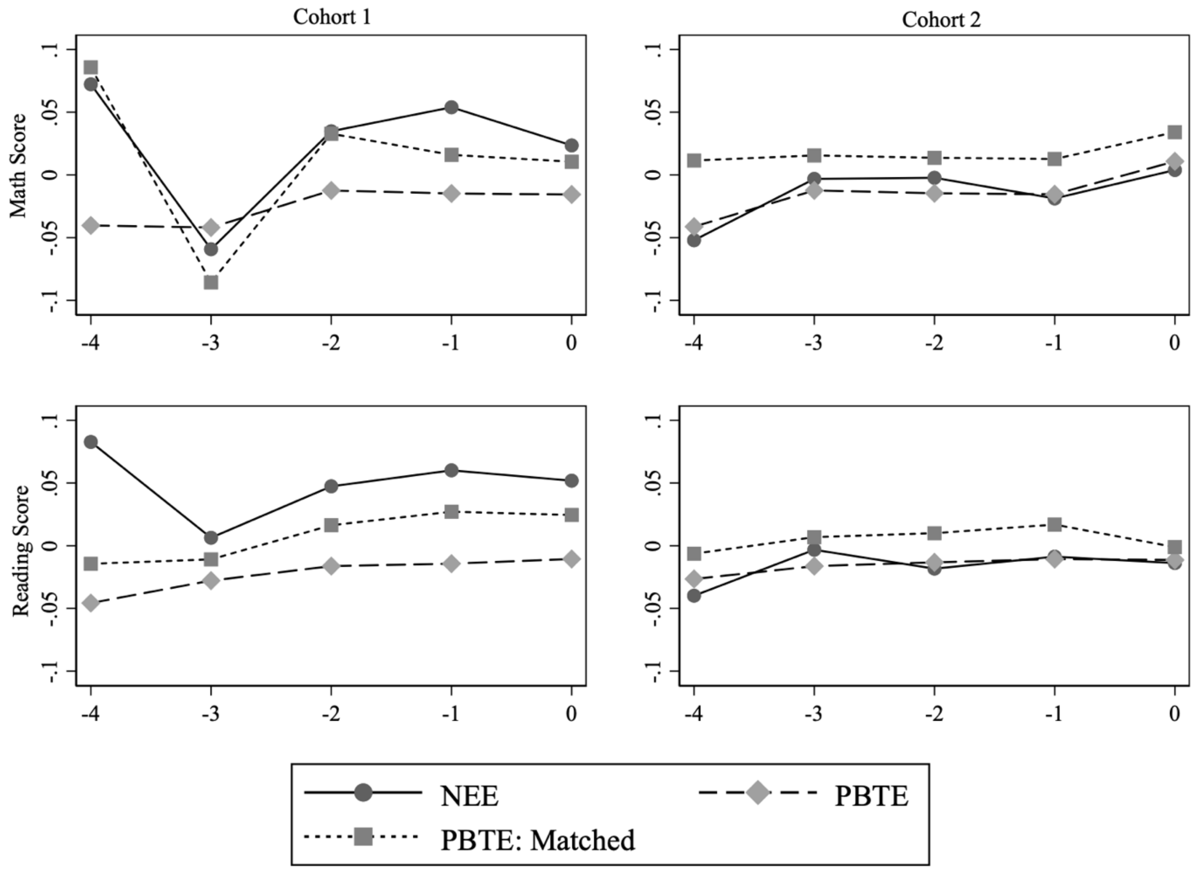
Taylor, E. S., & Tyler, J. H. (2012). The Effect of Evaluation on Teacher Performance. *American Economic Review*, *102*(7), 3628–3651. https://doi.org/10.1257/aer.102.7.3628

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.

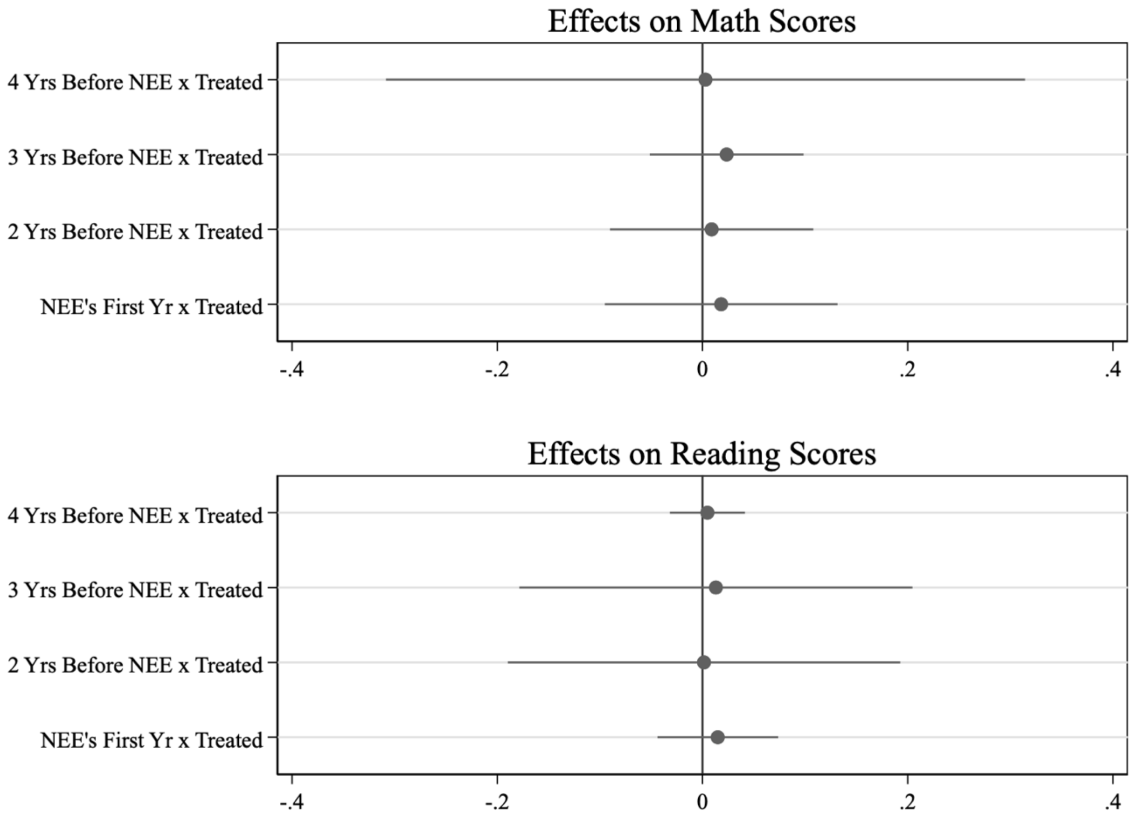
Woulfin, S. L., & Rigby, J. G. (2017). Coaching for Coherence: How Instructional Coaches Lead Change in the Evaluation Era. *Educational Researcher*, *46*(6), 323–328. https://doi.org/10.3102/0013189X17725525

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



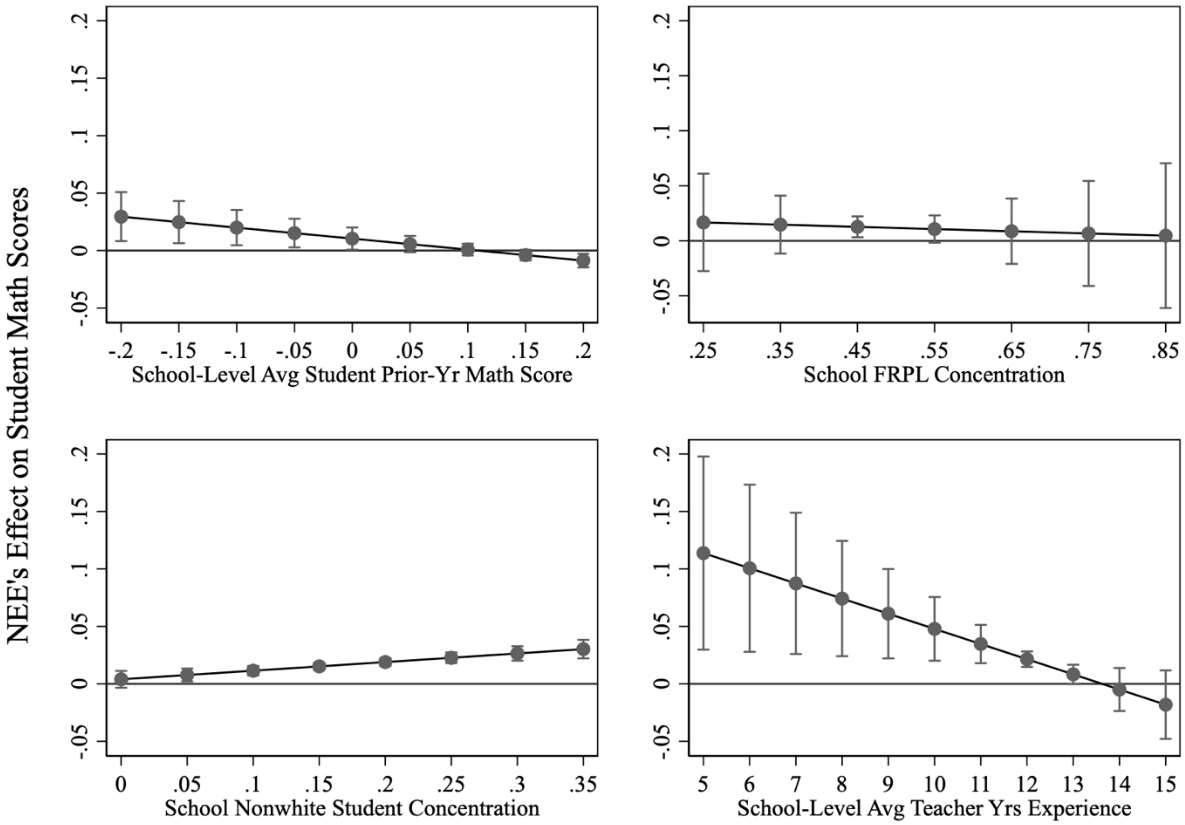
Notes: Each point represents average district-level average-student achievement scores; districts are the unit of analysis. Year 0 represents NEE's introduction. Top panels plot math scores, bottom panels plot reading scores; left panels plot Cohort 1 trends, right panels Cohort 2 trends.

Figure 2. Nonparametric Event Study Estimates



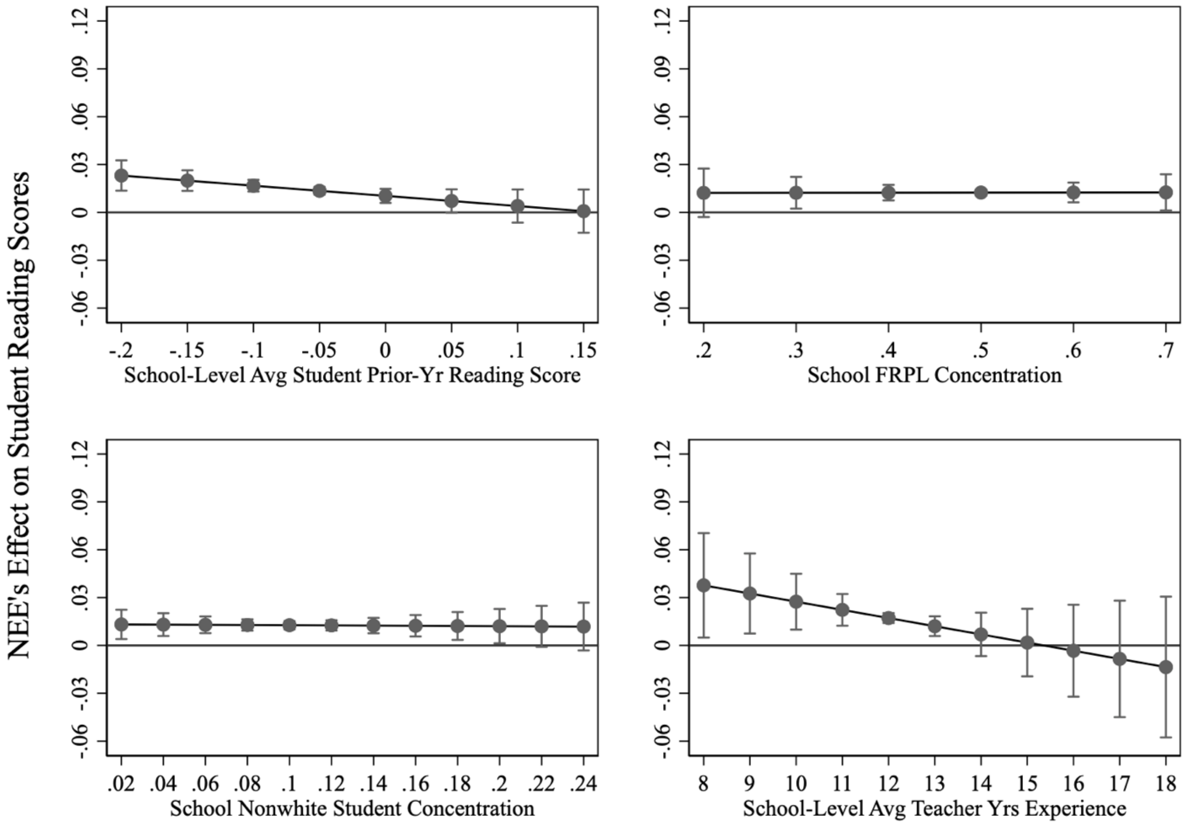
Notes: Point estimates and 95 percent confidence intervals. The top panel represents NEE's 'effects' on math scores relative to the 'effect' one year prior to NEE's introduction. The bottom panel represents analogous 'effects' on reading scores. Students are the unit of analysis. Models apply district-cohort 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. N(Math Student-Yrs) = 319096. N(Reading Student-Yrs) = 456232.

Figure 3. Moderating Effects of School Characteristics on NEE's Effect on Math Scores



Notes: Point estimates and 95 percent confidence intervals represent NEE's effect on math scores moderated by a linear school-level characteristic. Models interact treatment with a linear moderator, apply district-cohort fixed effects, year-cohort fixed effects, and control for urbanicity, student prior-year math score, district-level prior-year PPE, and the linear moderator. Standard errors multiway clustered by district, student, and cohort. N(Student-Yrs) = 319096.

Figure 4. Moderating Effects of School Characteristics on NEE's Effect on Reading Scores



Notes: Point estimates and 95 percent confidence intervals represent NEE's effect on reading scores moderated by a linear school-level characteristic. Models interact treatment with a linear moderator, apply district-cohort fixed effects, year-cohort fixed effects, and control for urbanicity, student prior-year reading score, district-level prior-year PPE, and the linear moderator. Standard errors multiway clustered by district, student, and cohort. N(Student-Yrs) = 456232.

Table 1. Features of PBTE and NEE Teacher Evaluation Systems

	Performance-Based Teacher Evaluation	Network for Educator Effectiveness
Timeline: Introduction and Retirement of System	1982-83 through 2012-2013	2011-12 until present
Observation Protocol	Performance rubric based on Missouri-specific standards for teaching	Similar to Danielson's Framework for Teaching, aligned to Missouri teaching standards
Grain Size of Observation: How many performance indicators (e.g., questioning, content knowledge, classroom management) are to be scored in an observation?	One, two, or six indicators	Three to five indicators
Integration with Professional Development Systems	No clear systematic integration	Online professional development library linked to the performance indicators in observation protocol
Sampling Procedure: Approximate Length of Observation	Unspecified	"Short" mini-observations.
Sampling Procedure: Frequency of Observations	Recommended new teachers receive one scheduled, two unscheduled for first three years. After third year one scheduled and one unscheduled. Tenured teachers observed only during formal evaluation year.	Recommend all teachers receive six to ten mini-observations each year.
Scoring Procedure: Is a final score produced after each observation? Is it a mean? Is a score determined holistically?	Holistically determined score using 3-point scale.	Score generated after each observation for each focal indicator on 5-point scale.
Observer Preparation/ Certification	No evidence of systematic preparation or credentialing system.	Annual and ongoing training to ensure reliable and accurate observation scoress, effective post-observation feedback conferences. Observers take a qualifying exam each summer.
Post-Conference Occurrence	After each observation	After each observation

Table 2. Descriptive Statistics

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

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

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

	I	II	III	IV	V	VI
Panel A. Math Scores						
Panel A1. Pooled Effects						
NEE	0.01	0.01	0.01	0.01	0.01	
	[-0.02,0.04]	[-0.05,0.07]	[-0.10, 0.13]	[-0.04, 0.06]	[-0.05, 0.07]	
Panel A2. Effects Moderated by Cohort						
NEE: Cohort 1						0.01
						[-0.03, 0.05]
NEE: Cohort 2						0.01
						[-0.02, 0.04]
N(Student-Yr)	319096	319096	319096	319096	319096	319096
Panel B. Reading Scores						
Panel B1. Pooled Effects						
NEE	0.01	0.01	0.01	0.01	0.01	
	[-0.00,0.03]	[-0.04, 0.06]	[-0.04, 0.06]	[-0.02, 0.05]	[-0.04, 0.06]	
Panel B2. Effects Moderated by Cohort						
NEE: Cohort 1						0.02
						[-0.05, 0.09]
NEE: Cohort 2						0.01
						[-0.03, 0.05]
N(Student-Yr)	456232	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			X
Year-Cohort FE	X	X	X			X

Notes: Point estimates and 95 percent confidence intervals in brackets 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 multiway clustered by district, student, and cohort.

Table 4. NEE's Total Effects on Cohort 1: One and Two Years of Implementation

Panel A. Math Scores		
NEE: Cohort 1 Year 1	0.01	[-0.02, 0.03]
NEE: Cohort 1 Year 2	0.01	[-0.02, 0.03]
N(Student-Yr)	127593	
Panel B. Reading Scores		
NEE: Cohort 1 Year 1	0.01	[-0.03, 0.06]
NEE: Cohort 1 Year 2	0.01	[-0.01, 0.03]
N(Student-Yr)	194166	

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

Table 5. Placebo Tests

	I	II	III	IV
Years Preceding NEE	t-1	t-2	t-3	t-4
Panel A. Math Scores				
NEE	0.001	0.02	-0.01	0.002
	[-0.06,0.06]	[-0.15,0.18]	[-0.28,0.26]	[-0.05,0.05]
N(Student-Yr)	319096	319096	319096	319096
Panel B. Reading Scores				
NEE	0.001	-0.005	0.01	-0.002
	[-0.15, 0.15]	[-0.18, 0.18]	[-0.24, 0.26]	[-0.02, 0.005]
N(Student-Yr)	456232	456232	456232	456232

Notes: Point estimates and 95 percent confidence intervals in brackets 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.

Table 6. Balance of Observables

Panel A. Student Characteristics	Math Students		Reading Students	
Female	< 0.01	[-0.05, 0.05]	0	[-0.07,0.06]
Nonwhite	< 0.01	[-0.09, 0.10]	0	[-0.03,0.03]
FRPL	< 0.01	[-0.05, 0.05]	0	[-0.06,0.06]
Prior-Year Achievement Score	0.01	[-0.02, 0.03]	-0.01	[-0.06,0.03]
Panel B. School Characteristics				
Concentration Female Students	< 0.01	[-0.01, 0.02]	0	[-0.04,0.04]
Concentration Nonwhite Students	< 0.01	[-0.01, 0.01]	0	[-0.05,0.04]
Concentration FRPL Students	< 0.01	[-0.02, 0.02]	0	[-0.06,0.06]
Avg Stdt Prior-Yr Ach Score	-0.01	[-0.15, 0.12]	-0.01	[-0.06,0.03]
Concentration Female Teachers	-0.01	[-0.05, 0.03]	0	[-0.09,0.08]
Concentration Nonwhite Teachers	< 0.01	[-0.01, 0.02]	0	[-0.03,0.03]
Concentration Adv Degrees	< 0.01	[-0.03, 0.03]	0	[-0.04,0.05]
Avg Tch Years of Experience	-0.14	[-1.94, 1.66]	0.01	[-2.14,2.15]
Panel C. District Characteristics				
Concentration Female Students	< 0.01	[-0.05, 0.04]	0	[-0.03,0.03]
Concentration Nonwhite Students	< 0.01	[-0.09, 0.10]	0	[-0.03,0.03]
Concentration FRPL Students	< 0.01	[-0.05, 0.04]	0	[-0.04,0.03]
Avg Stdt Prior-Yr Ach Score	< 0.01	[-0.11, 0.11]	-0.01	[-0.06,0.03]
Concentration Female Teachers	< 0.01	[-0.10, 0.09]	0	[-0.14,0.14]
Concentration Nonwhite Teachers	< 0.01	[-0.01, 0.01]	0	[-0.01,0.01]
Concentration Adv Degrees	0.01	[-0.00, 0.02]	0.01	[-0.02,0.03]
Avg Tch Years of Experience	0.04	[-1.17, 1.25]	0.09	[-0.50,0.68]
Prior-Year Per Pupil Expenditure	-33.37	[-2726.93, 2660.18]	77.39	[-2035.29,2190.08]
N(Student-Yr)	319602		456232	

Notes: Point estimates and 95 percent confidence intervals in brackets represent NEE's 'effect' on each covariate. 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.

Table 7. NEE's Effect on Teacher Mobility: Math Sample

	I	II	III	IV	V	VI
Panel A. Switched Districts						
Panel A1. Pooled Effects						
NEE	-0.03	-0.04	-0.03	-0.03	-0.03	
	[-0.11,0.04]	[-0.18,0.10]	[-0.13,0.07]	[-0.19,0.12]	[-0.19,0.12]	
Panel A2. Effects Moderated by Cohort						
NEE: Cohort 1						-0.05
						[-0.10,0.00]
NEE: Cohort 2						0.02
						[-0.04,0.08]
N(Teacher-Yr)	11748	11748	11748	11748	11748	11748
Panel B. Exited Teaching						
Panel B1. Pooled Effects						
NEE	0.00	0.00	0.00	0.00	0.01	
	[-0.04,0.04]	[-0.19,0.20]	[-0.45,0.45]	[-0.07,0.07]	[-0.16,0.18]	
Panel B2. Effects Moderated by Cohort						
NEE: Cohort 1						0.00
						[0.00,0.00]
NEE: Cohort 2						0.02
						[-0.11,0.15]
N(Teacher-Yr)	11748	11748	11748	11748	11748	11748
Controls		X			X	
District FE				X	X	
Year FE				X	X	
Cohort FE				X	X	
Controls-Cohort			X			
Dist-Cohort FE	X	X	X			X
Year-Cohort FE	X	X	X			X

Notes: Point estimates and 95 percent confidence intervals in brackets represent NEE's effect on teacher mobility. Panel A (B) coefficients represent the probability a teacher switches to a new district (exits the MO teacher labor market) instead of remaining in their school. All models control for urbanicity and district-level prior-year PPE. Standard errors multiway clustered by district, student, and cohort. The sample includes teachers from NEE districts and matched districts based on historical district-level average student math achievement scores only. Findings from the analogous sample of matched districts based on historical district-level average student reading achievement scores only is in Table C1. Estimates in Table C1 resemble estimates in Table 7.

Online Appendix A. Baseline Balance Tests

We test baseline balance using Equation A:

$$x_{isdt} = \delta NEE_{dt} + \beta_1 y_{isd(t-1)} + \beta_2 PPE_{d(t-1)} + \beta_3 Rural_{dt} + \Delta_{dc} + \Phi_{tc} + e_{isdtc} \quad (A),$$

where x_{isdt} represents student prior-year achievement scores, school- or district-level average student prior-year achievement scores, or prior-year PPE, none of which NEE's introduction can affect genuinely as these four 'outcomes' were measured prior to NEE's launch. A statistically significant δ in Equation A would imply that matched PBTE districts, or the schools and students within, were not chosen in a way that mimics 'randomization.' Other terms refer to the same quantities as Equation 1.

Online Appendix B. Coarsened Exact Matching Results

Table B1. 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 B2. 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 B1 notes.

Online Appendix C. Teacher Mobility Using Reading Sample

Table C1. NEE's Effect on Teacher Mobility: Reading Sample

	I	II	III	IV	V	VI
Panel A. Switched Districts						
Panel A1. Pooled Effects						
NEE	-0.03*	-0.04	-0.04	-0.03	-0.04	
	[-0.06,-0.00]	[-0.11,0.04]	[-0.16,0.09]	[-0.14,0.07]	[-0.13,0.05]	
Panel A2. Effects Moderated by Cohort						
NEE: Cohort 1						-0.03*
						[-0.04,-0.02]
NEE: Cohort 2						0.00
						[-0.04,0.05]
N(Teacher-Yr)	17781	17781	17781	17781	17781	17781
Panel B. Exited Teaching						
Panel B1. Pooled Effects						
NEE	0.00	-0.01	-0.01	0.00	-0.01	
	[-0.06,0.05]	[-0.16,0.13]	[-0.27,0.25]	[-0.08,0.07]	[-0.15,0.13]	
Panel B2. Effects Moderated by Cohort						
NEE: Cohort 1						0.00
						[0.00,0.00]
NEE: Cohort 2						0.00
						[-0.04,0.03]
N(Teacher-Yr)	17781	17781	17781	17781	17781	17781
Controls		X			X	
District FE				X	X	
Year FE				X	X	
Cohort FE				X	X	

Controls-Cohort			X	
Dist-Cohort FE	X	X	X	X
Year-Cohort FE	X	X	X	X

Notes: Point estimates and 95 percent confidence intervals in brackets represent NEE's effect on teacher mobility. Panel A (B) coefficients represent the probability a teacher switches to a new district (exits the MO teacher labor market) instead of remaining in their school. All models control for urbanicity and district-level prior-year PPE. Standard errors multiway clustered by district, student, and cohort. The sample includes teachers from NEE districts and matched districts based on historical district-level average student reading achievement scores only.