education policy analysis archives

A peer-reviewed, independent,
open access, multilingual journal

Arizona State University

Volume 34 Number 42 May 5, 2026 ISSN 1068-2341

Reformed Teacher Evaluation in Rural Missouri: Main and Moderated Relationships with Student Achievement and Relationships-to-Expenditure Ratios

Seth B. Hunter
George Mason University
United States
&
Katherine M. Bowser
National Council on Teacher Quality
United States

Citation: Hunter, S. B., & Bowser, K. M. (2026). Reformed teacher evaluation in rural Missouri: Main and moderated relationships with student achievement and relationships-to-expenditure ratios. *Education Policy Analysis Archives, 34*(42). https://doi.org/10.14507/epaa.34.9248

Abstract: We extend rural educator labor market research by estimating a reformed teacher evaluation system's relationships with student achievement, identifying the settings with positive relationships, and incorporating evaluation expenditures. That the literature omits these contributions is concerning, as research implies it hinders evidence-based policymaking for rural districts, which outnumber urban districts in the USA. We apply a difference-in-differences framework to rural Missouri administrative data. Missouri districts could design and maintain reformed systems or outsource these tasks for a small fee to organizations like the Network for Educator Effectiveness (NEE), an evaluation system created for rural users. NEE does not affect student achievement on average but may improve math and possibly

Journal website: http://epaa.asu.edu/ojs/
Facebook: /EPAAAAPE
Bluesky: @epaa.bsky.social

Manuscript received: 1/4/2025
Revisions received: 12/11/2025
Accepted: 9/2/2026

reading achievement in rural schools where the average student's prior-year achievement score is below the state average or the average teacher's years of experience are below the state average.
Keywords: rural education; education policy; school/teacher effectiveness; quasi-experiments

Evaluación docente reformada en zonas rurales de Missouri: Relaciones principales y moderadas con el rendimiento estudiantil y y su vínculo con el gasto educativo
Resumen: Ampliamos la investigación sobre el mercado laboral docente en contextos rurales al estimar las relaciones de un sistema reformado de evaluación docente con el rendimiento estudiantil, identificar los contextos en los que dichas relaciones son positivas e incorporar los gastos asociados a la evaluación. La ausencia de estas contribuciones en la literatura resulta preocupante, ya que la investigación sugiere que ello obstaculiza la formulación de políticas basadas en evidencia para los distritos rurales, que superan en número a los distritos urbanos en los Estados Unidos. Aplicamos un marco de diferencias en diferencias a datos administrativos de zonas rurales de Missouri. Los distritos de Missouri podían diseñar y mantener sistemas reformados o externalizar estas tareas por una tarifa reducida a organizaciones como la Network for Educator Effectiveness (NEE), un sistema de evaluación creado para contextos rurales. En promedio, el NEE no afecta el rendimiento estudiantil, pero puede mejorar el rendimiento en matemáticas y posiblemente en lectura en escuelas rurales donde el puntaje promedio de logro del estudiantado en el año previo está por debajo del promedio estatal o donde los años promedio de experiencia docente están por debajo del promedio estatal.
Palabras clave: educación rural; política educativa; efectividad escolar/docente; cuasi-experimentos

Avaliação docente reformada em áreas rurais de Missouri: Relações principais e moderadas com o desempenho dos estudantes e sua relação com os gastos educacionais
Resumo: Ampliamos a pesquisa sobre o mercado de trabalho docente em contextos rurais ao estimar as relações de um sistema reformado de avaliação docente com o desempenho dos estudantes, identificar os contextos em que essas relações são positivas e incorporar os gastos associados à avaliação. A ausência dessas contribuições na literatura é preocupante, uma vez que pesquisas indicam que isso dificulta a formulação de políticas baseadas em evidências para distritos rurais, que superam numericamente os distritos urbanos nos Estados Unidos. Aplicamos um modelo de diferenças em diferenças a dados administrativos de áreas rurais de Missouri. Os distritos de Missouri puderam desenvolver e manter sistemas reformados ou terceirizar essas tarefas por uma pequena taxa a organizações como a Network for Educator Effectiveness (NEE), um sistema de avaliação criado para contextos rurais. Em média, o NEE não afeta o desempenho dos estudantes, mas pode melhorar o desempenho em matemática e possivelmente em leitura em escolas rurais onde a média de desempenho dos estudantes no ano anterior está abaixo da média estadual ou onde a média de anos de experiência docente também está abaixo da média estadual.
Palavras-chave: educação rural; política educacional; efetividade escolar/docente; quase-experimentos

Reformed Teacher Evaluation in Rural Missouri: Main and Moderated Relationships with Student Achievement and Relationships-to-Expenditure Ratios

Rural settings are woefully understudied in education policy research (Tran, 2023). This is concerning given more than 50% of school districts in the United States are in rural settings (School Superintendents, 2017) and policymakers prefer research conducted in settings similar to their own when making data-driven decisions (Nakajima, 2021). Taken together, most local education policymakers (i.e., district leaders in rural settings) are left out of important research on educational policy and practice. In fact, from 2004–2014, 64 journal articles in the then-top five ranked education journals included "urban" in the title or abstract, while only 5 included "rural" (Schafft & Biddle, 2014). This "peripheralization" (p. 138) of rural educational research is detrimental to rural communities and economies broadly (Schafft, 2016) and leads to "the continual undercutting of support for their teachers" (Tran, 2023, p. 238). This undercutting of support is particularly true regarding teacher evaluation research, which is meant to inform policymakers and practitioners, but only a few urban-centric studies have informed national narratives about teacher evaluation reforms despite concerns regarding generalizability to rural settings.

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

Despite the widespread adoption and importance of evaluation reforms, rigorous quantitative research examining evaluations' effects on student achievement is thin.[1] Most of what we have learned comes from a few urban centers (Dee & Wyckoff, 2015; Steinberg & Sartain, 2015; Taylor & Tyler, 2012) and one U.S. national study (Bleiberg et al., 2024). Findings suggest that evaluations' effects on student achievement range from near-zero to substantially positive, though most effects are near-zero. Importantly, two studies also examine school characteristics moderating these effects, enabling targeted policy implications; one finds that the magnitude of positive effects increases with teacher years of experience (Taylor & Tyler, 2012), while another concludes that evaluations' positive effects increase with school-level student economic advantage and prior-year achievement scores (Steinberg & Sartain, 2015); both suggesting that evaluation reforms may exacerbate inequality.

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

However, no quasi-experimental research focuses on the costs or effects of teacher evaluation on student achievement in rural settings. There are at least two reasons why urban-centric research may not generalize to rural teacher evaluation. First, rural districts may have inequitable access to the resources (e.g., money, time, personnel) needed to implement reformed teacher evaluation systems with fidelity. Indeed, the initial and ongoing research, development, and support for the implementation of reformed teacher evaluation systems, including teacher performance rubric and measure design, professional development for principals (teacher evaluation's primary implementers), and maintenance of a performance data management system, requires education agency time, capacity, and resources that may be prohibitive for rural districts (Chambers et al., 2013; Gilles, 2017; Stecher et al., 2018). While some urban districts have the capacity (e.g., specialized staff) to build and maintain evaluation systems, many rural districts have outsourced this work. In Missouri—the setting of this study—over 300 districts outsource their evaluation processes and systems to a third-party organization for a fee. Presumably, this fee is less than the total cost of what a district would otherwise spend to develop and maintain its evaluation system; a presumption we examine in detail.

Second, the unique challenges of rural teacher labor markets may dissuade principals from dismissing underperforming teachers, suppressing a core mechanism by which teacher evaluation aims to improve the distribution of teacher effectiveness (Rodriguez et al., 2020). For example, if principals avoid providing teachers with critical performance feedback following classroom observations to ensure they remain in their school, principals are foregoing practically large feedback-induced teaching improvements (Hunter & Springer, 2022; Hunter & Steinberg, 2022). These conditions suggest that urbanicity and its correlates may moderate evaluations' effects on student outcomes. Additionally, no rurally focused study links these effects to expenditures, information policymakers need to make informed decisions.

Ultimately, there is insufficient evidence to reach defensible conclusions about the costs and effects of teacher evaluation reforms on rural student achievement scores and even less evidence regarding the conditions in which these reforms improve rural student outcomes. We address these gaps by answering the following research questions:

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

We investigate a unique teacher evaluation system, the Network for Educator Effectiveness (NEE). While prior work focuses on teacher evaluation systems managed by state departments of education or district central offices,[2] an independent center at the University of Missouri manages NEE. Districts join NEE voluntarily and pay an annual fee to cover ongoing operational costs. Despite the availability of a state-designed evaluation system without fees and the option for districts to design their system, the number of districts choosing NEE has grown from 6 of Missouri's 500+ districts in 2011–12 to nearly 300 Missouri districts, 20 in Nebraska, 4 in Kansas, and 1 each in Oregon and Illinois, as of the 2025–26 school year, underscoring the relevance of examining this system.

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

While any district can become a NEE member, NEE was designed with rural districts in mind. The content of NEE rubrics and the teacher indicators measured do not differ substantively from other evaluation systems, rather NEE addresses the two conditions we argue may moderate teacher evaluation's effects by urbanicity or its correlates. It helps districts cope with the human, financial, and physical resource constraints in rural settings by developing and maintaining a theoretically sound teacher evaluation system for a remarkably low cost due to economies of scale and location inside a public university system, which offsets some of NEE's operational costs. NEE is also sensitive to the reality of rural teacher labor markets and eschews teacher evaluation's accountability mechanism completely. Consequently, NEE focuses on teacher development exclusively and expands the developmental capacities of rural central offices. We expand on why the focus on teacher evaluation's developmental mechanism is particularly salient to rural settings in our theory of action section.

We compare NEE's fee per student to other documented teacher evaluation costs and apply quasi-experimental methods to five years of panel data to estimate NEE's average relationships with rural student achievement. We also estimate heterogeneous relationships in policy-relevant school settings. Our average quasi-experimental estimates are precisely estimated, negligibly positive, and statistically insignificant, resembling findings from other settings. However, the data repeatedly suggest that NEE student achievement scores increase in schools where the average student's prior-year achievement score and the average teacher's years of experience are below state averages, contrary to the findings in non-rural settings. District expenditures suggest that NEE's annual costs to districts are among the cheapest of any documented system; consequently, relationships-to-expenditure ratios in settings with positive estimates are notable for these specific subgroups. Our findings suggest that rurally concerned policymakers might consider piloting teacher evaluation systems resembling NEE in disadvantaged schools, recognizing that observed benefits are modest and context specific.

Background

Theory of Action Framing Reformed Teacher Evaluation Systems

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

The developmental components of evaluation reforms might also improve teaching quality. Observation conferences can provide teachers with performance-enhancing strategies. As reformed systems include higher frequencies of observations and post-observation feedback conferences (Steinberg & Donaldson, 2016), teachers effectively receive higher dosages of performance feedback. Notably, the feedback itself may not improve teaching directly (Cherasaro et al., 2016; Hunter & Springer, 2022; Ilgen et al., 1979; Murphy & Cleveland, 1995). Instead, feedback may lead

teachers to PD opportunities tailored to observation-identified areas of weakness (e.g., targeted coaching; Donaldson, 2021), underscoring the importance of linkages between evaluation and PD systems (Kraft & Gilmour, 2017; Weisberg et al., 2009). Ultimately, evaluation as a developmental tool theoretically depends on feedback, pointing towards the significance of evaluators' observation and feedback skills (Hunter & Steinberg, 2022; Kimball & Milanowski, 2009).

Accountability and developmental mechanisms may operate differently in rural districts, such that rural settings are particularly well-suited for developmentally focused teacher evaluation systems. Due to a smaller labor pool and lower salaries, rural teacher labor markets are less elastic than non-rural markets and rural schools face unique challenges in recruiting and retaining teachers compared to their urban peers (Nguyen et al., 2020). For example, states with low population density (including Missouri, our study setting) experience higher rates of novice teacher turnover in rural communities than in urban centers, effectively placing rural schools in a constant state of onboarding and development (Nguyen et al., 2020). Further, Ingersoll and Tran (2023) identify a "revolving door of teachers" in rural districts that is primarily driven by preretirement turnover, with over 60% of teachers who left their position citing dissatisfaction with administration as one of the most important reasons in their decision to leave (p. 413). Importantly, prior literature has identified strong community ties as an important advantage to rural teaching that can improve retention (Tran et al., 2020). Due to these characteristics, we theorize three reasons a developmental teacher evaluation system may be a promising strategy for rural districts.

First, the central pillar of the accountability mechanism is the assumption that when an ineffective teacher exits (voluntarily or involuntarily), they are replaced with a more effective teacher (Donaldson, 2021). However, the limited labor supply in rural areas means that a more effective replacement is not guaranteed and thus may hinder principals' willingness to dismiss ineffective teachers. Prior work supports this theory; Rodriguez et al. (2020) examined the introduction of a reformed evaluation system in Tennessee and found it to be associated with increased *retention* of *more* effective teachers and increased *turnover* of *less* effective teachers, consistent with conceptualizations of accountability mechanisms. Crucially, findings reflected differential mobility patterns in urban districts, but not in rural districts. The heterogeneity in results by urbanicity may be due to school leaders' hesitancy to dismiss underperforming teachers due to the challenges of finding replacements (Kraft & Gilmour, 2016; Rodriguez, 2020). It follows that a developmental teacher evaluation system that prioritizes supporting teachers by providing the feedback and resources needed for professional growth is particularly important in rural settings.

Second, to the extent that administrators provide helpful feedback and invest time in teacher improvement through developmental evaluation, teachers may perceive this strong administrative support as an improvement to their working conditions and/or job satisfaction. Superintendents, principals, and teachers who are aspiring administrators in rural districts agree with this theory. In a qualitative study, they identified a strong commitment to providing structural support to teachers, like providing feedback and professional development opportunities—key characteristics of a developmental evaluation system—is crucial to engaging new teachers (Frahm & Cianca, 2021). This administrative support is particularly salient in rural districts as dissatisfaction with administration was one of the most important factors in rural teachers' decision to leave their position (Ingersoll & Tran, 2023).

Finally, developmentally focused evaluation systems provide structured time for one-on-one conversations between teachers and principals. Rural leaders identified the interactions resulting from observations as opportunities to connect with teachers and "to deepen existing relationships and prevent isolation" (Frahm & Cianca, 2021, p. 7). This is particularly advantageous in rural settings where community and relationships are a driver of teacher retention and isolation is a driver of turnover (Nguyen et al., 2020; Tran et al., 2020).

Importantly, the effectiveness of a developmental teacher evaluation system to improve teacher quality (not just staffing challenges) may still be a challenge in rural settings. Prior work finds that teacher performance improves when observers provide more critical feedback; however, receiving more critical feedback may (unintentionally) result in teacher turnover as teachers leave critical feedback settings and move into positive feedback schools (Feng, 2010; Hunter & Springer, 2022; Hunter & Steinberg, 2022). Thus, rural school leaders may forego critical feedback that could improve teacher performance to avoid the challenges of finding replacements. Qualitative work supports this theory as principals implementing a newly reformed evaluation system noted that they become lenient with their feedback as a relationship-managing strategy (Derrington, 2014).

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

Related Studies Regarding Teacher Evaluation Effects

We focus on the plausibly causal effects of introducing reformed teacher evaluation systems on student achievement scores, which only a few predominantly urban studies examine.[3] In a randomized control trial, Steinberg and Sartain (2015) estimated the effects of a reformed teacher evaluation pilot, the Excellence in Teaching Project (EITP). EITP, a low-stakes system, was implemented across two cohorts of elementary schools in Chicago Public Schools. While analyses of student math scores did not detect effects, student reading scores increased significantly, although these effects were concentrated in the first EITP cohort. This is the only study we know of that estimates the heterogeneous effects of school-level characteristics; advantaged schools (i.e., higher-performing and lower-poverty) benefited more than disadvantaged schools. There was no evidence of heterogeneity by school-level shares of student race or average teacher years of experience.

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

A recent nationwide study using data from the Stanford Education Data Archive found that reforms had, on average, no effect on student achievement (Bleiberg et al., 2024). The authors also examined heterogeneity across system design features but still estimated precise null effects. Notably, the authors hypothesized that ineffective implementation explains the lack of student achievement gains.

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

Cost Studies

Two studies examine the per pupil expenditures associated with teacher evaluation systems using data from the Intensive Partnerships for Effective Teaching (IP) initiative. Seven districts—three school districts and four charter management organizations—implemented teacher evaluation systems from 2009–10 through 2014–15. The systems examined included reforms shared by NEE, EITP, and the Cincinnati systems. The overall per-pupil expenditures ranged from $868 to $3,541 (Stecher et al., 2018). For five of the seven districts, one-time bonuses and permanent salary increases tied to teacher evaluation were the largest sources of expenditures. Principal and teacher PD were the largest source of expenditures for the remaining two districts. Additionally, Stecher et al. (2018) determined that accounting for the time teachers and school leaders spent on evaluation would add $200 per pupil, on average.

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

CBO found that the per pupil yearly expenditures ranged from $8 to $118 across system components over the first three years of implementation. Each district spent the highest proportion (47%–87%) of funds on the teacher observation component, though the subdomains driving these expenditures varied across districts. One district spent a great deal on 'peer, mentor, and external evaluators' while the remaining districts relied almost exclusively on principals and assistant principals to conduct observations—as do NEE districts. The second most expensive subdomain for each district was 'management and communications,' followed by 'technology and data.' Importantly, CBO likely underestimated the total costs for two reasons. First, the expenditures do not include the time spent by district personnel (who were not hired explicitly to manage the reformed system). Second, the districts examined began prepping for the reformed evaluation systems before the study period; thus, some costs may have occurred before the study period.

Study Context

In the early 2010s, researchers at the University of Missouri's College of Education developed NEE with substantial input from recently retired PreK-12 rural principals and district administrators. NEE recruited six districts into Cohort 1 to pilot the system in 2011–12 and launched training during the summer of 2011. Notably, Cohort 1 districts did not know about the summer 2011 launch until spring 2011, and districts did not receive their 2010–11 student achievement scores until after recruitment. These conditions mitigate concerns about the endogenous timing of NEE's adoption and anticipatory effects, particularly those arising from student achievement scores and their correlates. NEE recruited 26 districts into Cohort 2 for a

2012–13 launch and trained those districts in the summer of 2012 before participants received 2011–12 achievement scores. Recruitment in both cohorts was based entirely on district urbanicity and NEE developers' professional networks with rural superintendents. All recruited districts joined, and superintendents treated NEE as a district-level policy that they expected all their schools to implement.

NEE Design Element 1: Observation Rubrics and Goal Setting

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

Setting individual performance goals using performance rubrics is a theoretically essential component of any evaluation system (Choi & Johnson, 2022; Locke & Latham, 2002); indeed, recent work suggests rubric-based goal setting may be one of the main mechanisms by which evaluation improves early-career teacher performance (Hunter & Springer, 2022). NEE teachers actively engage in goal-setting processes with school administrators to select annual teacher performance goals tied to 3 of the 39 NEE rubric indicators.

NEE Design Element 2: Observation Frequency and Conferences

Theoretically, classroom observations are the linchpin of evaluation for development as they can include teaching performance assessment, goal-setting, performance-enhancing feedback, and improvement plans (Donaldson, 2021). Empirically, the effects of more observations on student achievement vary by context. Research from DC's high-stakes evaluation system suggests that the marginal observation improves teaching and student achievement (Phipps, 2022; Phipps & Wiseman, 2021). However, larger-scale research from more typical low-stakes settings finds no effects on student achievement (Hunter & Kho, 2025). Combined with the findings from Hunter and Steinberg (2022), we interpret the evidence and theory to mean that growth is more likely when teachers are observed more frequently.

NEE recommends that every teacher receive six to ten mini-observations per year. In theory, shorter frequent observations permit principals to see a wider range of lesson-sensitive instructional skills than fewer, longer observations and are more conducive to principals' schedules—particularly in rural districts that are less likely to have specialized staff dedicated to conducting observations. These theories have been reflected in interview studies with principals who perform teacher observations (Carraway & Young, 2015; Ovando & Ramirez, 2007). During the study period, we do not know how many observations were received per teacher; however, prior work from other settings finds that principals typically conduct fewer observations than teachers are assigned (Hunter & Ege, 2021; Hunter & Kho, 2025; Kraft & Gilmour, 2016). Indeed, NEE observation data collected after the period examined by our study reveals that the typical teacher received four yearly observations (Hunter & Steinberg, 2022). Though this is below the NEE recommendation, it is an unusually high number of observations relative to other systems and may result in greater gains in student achievement (Hunter, 2020; National Council on Teacher Quality, 2019).

NEE Design Element 3: Observer Preparation and Certification

NEE evaluators receive annual and ongoing training and support to promote reliable and accurate scoring. Evaluators also receive training about how to provide performance feedback

effectively. Training also focuses on collaboration with teachers directly and supporting teacher collaboration with other personnel (e.g., peer mentoring) to improve observation-identified areas for growth. Following training, prospective evaluators must pass an exam each summer to receive certification to conduct formal observations. Theoretically, design element three should also increase the odds that NEE improves student achievement.

NEE Fees

NEE charges districts an average of $3 per student to cover its operational costs. We compare NEE's fee to the per pupil expenditures detailed in CBO. We use costs reported in CBO due to the detailed disaggregation of expense categories. Further, CBO reports districts' costs in the early years of policy adoption, like the timeframe studied herein. In NEE's first two years, the $3 per pupil fee included access to NEE's observation scoring rubric, the NEE Data Tool (a centrally managed performance data system), evaluator (school administrator) initial certification and yearly recertification training sessions, ongoing training for educators, webinars, and technical support via the NEE Help Desk.

Data and Methods

This study uses Grades 3–8 statewide administrative data from Missouri's Department of Elementary and Secondary Education (DESE), NEE-supplied lists of its first two cohorts, and National Center for Education Statistics (NCES) urbanicity and per-pupil expenditures (PPE) from 2007–08 through 2012–13. DESE allows the linkage of schools to districts, students to schools, and teachers to schools. Student administrative data includes race, gender, FRPL, and achievement scores, while teacher data includes race, gender, education level, and years of experience. As NEE is fee-based and designed for rural districts, we control for urbanicity and PPE via NCES data. NEE adoption is a district-level policy; thus, our independent variable is at the district-by-year level. Our outcome variable is student-by-year math and reading achievement. While the number of NEE districts is small, our sample size is sufficiently large. Cohort 1 included 24 schools enrolling approximately 5,000 students, and Cohort 2 included 71 schools enrolling approximately 10,000 students.[4]

Cost Analysis

To answer our first research question, we use the disaggregated cost analysis of CBO to construct relevant comparisons to NEE's fee. First, we align the components defined in CBO with NEE's products and services. Then, we identify conservative to liberal ranges of ongoing costs[5] that a district would have had to spend to implement a reformed teacher evaluation system.

Quasi-Experimental Analysis

NEE was not assigned to districts randomly; however, we use our strong knowledge of the selection process and NEE's discontinuous rollout to account for plausibly concerning confounders regarding the relationship between NEE's introduction on changes in student achievement. Although some recent research eschews generalized difference-in-difference designs (DDs) with two-way fixed effects due to concerns about heterogeneity arising from differential lengths of treatment exposure (Callaway & Sant'Anna, 2021; Goodman-Bacon, 2021), these concerns are

[4] For context, the data in Steinberg and Sartain (2015) included 44 schools in its first cohort and 48 in the second, and Taylor and Tyler (2012) included a rough total of 3,600 treated and untreated students.
[5] For completeness, we complete a similar comparison for start-up (i.e., one-time) costs.

unwarranted in our case for two reasons. First, we estimate cohort-specific relationships one year after treatment to examine cohort heterogeneity and one-year-after relationships pooled across cohorts. Second, when we estimate relationships for NEE's first cohort only, which was exposed to NEE for two years, we estimate cohort-by-year specific relationships.

We apply generalized DD designs, which assume that the changes in achievement scores among the treated would have been statistically equivalent to achievement score changes for the comparison group had the treated not been exposed to NEE. This supposition, the parallel trends assumption, is violated if time-varying omitted variables (OVs) affect the treated and comparison groups differentially and substantively in the pre- or post-treatment periods. While observed achievement scores allow us to estimate if they (and their correlates) changed in treated versus comparison groups differentially and substantively in the pre-treatment period, we do not observe post-treatment counterfactuals. However, limiting our analyses to two years after treatment, at most, decreases the chance that unobserved concurrent or post-exposure alternative treatments (i.e., OVs) affected student achievement trends in NEE districts but not comparison groups. Moreover, this short time frame, combined with the relatively small number of districts in Cohorts 1 and 2, means NEE leaders were aware of alternative treatments in their districts. Over two interviews, NEE leaders emphatically reported that NEE districts did not receive alternative treatments, bolstering the plausibility of parallel trends post-treatment.

Selection

NEE leaders also repeatedly stated that districts were recruited on two conditions: if the district was rural and its superintendent had a strong professional relationship with NEE leaders. NEE leaders also noted stronger ties with the superintendents in NEE Cohort 1 than Cohort 2. We observe urbanicity but not NEE's professional ties with superintendents; however, we argue that it is implausible for the professional ties between superintendents and NEE leaders to represent an OV capable of undoing our inferences. We accept that NEE leaders' relationships with superintendents might have been influenced by their perceptions of superintendent effectiveness, but we argue that selection on perceived superintendent effectiveness is not concerning. NEE would have needed to select on achievement-related superintendent effectiveness (instead of, for example, political acumen) to raise the possibility that treated achievement scores changed because of superintendent effectiveness. Not only did NEE leaders emphatically deny that student achievement affected their relationships with superintendents, but research concerning superintendent effects on student achievement suggests that such effects are relatively small and, therefore, unlikely to threaten our inferences (Schwartz et al., 2023).

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

Matched and Stacked Generalized Difference in Differences

We define treatment as exposure to NEE implementation, and the comparison group consists of students and schools in districts that did not implement NEE. We apply a matching procedure to strengthen the DD. We identify the non-NEE districts resembling NEE regarding PPE in the year before treatment and district-level average student achievement scores one, two,

three, and four years before treatment. We match on historical achievement trends and prior-year PPE in case they affected selection into a fee-based teacher evaluation system aiming to improve student achievement, despite the emphatic accounts otherwise from those who worked with NEE districts closely over several years.

Nonetheless, we use coarsened exact matching (CEM) per Sturge's Rule, in which districts are the unit of analysis since selection was at that level. Matching occurs by cohort; the pool of potential matches for Cohort 1 includes all rural districts that did not implement NEE through 2011–12, the year Cohort 1 launched. Districts that implemented NEE in 2012–13 were also in Cohort 1's pool of potential matches for 2011–12; Cohort 2 is omitted from the pool of potential matches for 2012–13. Cohort 2's matching procedure is analogous to Cohort 1's, except that the pool of potential matches includes all rural districts that did not use NEE through 2012–13. CEM matches on five variables: district-level PPE from the year before treatment and district-level average student achievement scores one, two, three, and four years before treatment.

Matching on district-level average student achievement scores one year before treatment, and the timing of NEE's recruitment bolsters the parallel trends assumption considerably. Student achievement scores from the spring before NEE's launch capture variation in all (unobserved) factors that determined those scores up to the point of assessment (Bacher-Hicks & Koedel, 2023; Cowan et al., 2022). Notably, NEE recruitment (selection) occurred before achievement testing in the spring before launch. These conditions suggest that matching on prior-year scores alone effectively controls for the OVs affecting selection and prior-year achievement. While our model does not control for factors affecting selection independent of prior-year achievement scores, such factors could only threaten the parallel trends assumption if they determined selection and post-treatment achievement scores but not pre-treatment scores from any of the four prior years; we assume such factors are unlikely.

After identifying district matches for Cohorts 1 and 2, matched data are returned to the student level and stacked; Cohort 1 and its matches are stacked onto the data for Cohort 2 and its matches, yielding a student-by-year-by-cohort dataset. Years within each cohort/ stack are centered on NEE's introduction year (e.g., Cohort/ Stack 1 year 0 corresponds with 2011–12) and range from -4 to 0. Following Gormley and Matsa (2011), we apply a generalized DD model to stacked data using Equation 1:

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

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

Sensitivity Tests

Our sensitivity tests begin by re-applying Equation 1 with an augmented set of control variables. If adding augmented controls to Equation 1 results in substantively different results, it

could indicate a violation of the parallel trends assumption. While we can control for observable differences, the sensitivity of estimates generated by Equation 1 to the augmented model would raise concerns about unobserved OVs. The augmented controls include student race, gender, FRPL, and the proportion of students in a school and district by race, gender, and FRPL; the concentration of teachers in a school and district by race, gender, education level, and average teacher years of experience; and school- and district-level average student prior-year achievement scores. We also estimate versions of Equation 1 using i) the canonical district FE and year FE, ii) district FE, year FE, and augmented controls, and iii) district-cohort FE, year-cohort FE, and cohort-specific augmented controls.

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

Heterogeneity

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

Parallel Trends Test and Event Study Analysis

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

$$y_{isdtc} = \delta_{-4}NEE_{dt} + \delta_{-3}NEE_{dt} + \delta_{-2}NEE_{dt} + \delta_{0}NEE_{dt} + \beta_1 y_{isd(t-1)} + \beta_2 PPE_{d(t-1)} + \tau_{dc} + \pi_{tc} + e_{isdtc} \quad (2)$$

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

Placebo Tests and Relationships with Covariates

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

We have no reason to believe that NEE's introduction should affect the observable compositions of NEE districts or any other covariate. Indeed, if treatment affects covariates, it could suggest alternative treatments. Appendix A describes the baseline balance tests in detail.

Relationships Over First Two Years: Cohort 1

Although the study's primary purpose is to identify relationships after one year of implementation, Cohort 1's data permit estimating NEE's relationships one and two years after introduction. We only retain Cohort 1 and its matched comparison group to estimate these dynamic relationships. 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, we apply district FE and year FE. We estimate dynamic post-intervention relationships by adding an interaction to Equation 1, interacting NEE_{dt} with an indicator marking if the records came from 2012–13.

Findings

Cost Comparison

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

Table 1

CBO Ongoing Yearly and Total Per Pupil Expenditure Estimates

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

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

As discussed above, CBO identified five teacher observation cost subdomains: i) design and implementation; ii) peer, mentor, and external evaluators; iii) management and communications; iv) technology and data systems; and v) other. We argue that CBO's (i), (iii), and (iv) subdomains clearly connect to NEE's services (e.g., observation rubric and related resources, observer training and calibration, a data management system, and technical support). Although we argue that NEE did not include expenditures concerning (ii) and (v), we add these to the upper bound costs for reference. Definitions of (i) – (v) and rationale for NEE's aligned services are detailed in Online Appendix Table B1. Finally, we adjust CBO costs to 2012 real dollars as NEE was launched in 2011–12.

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

Pre-Matched Descriptive Statistics

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

Matching Results for DD Design

As the validity of our strategy does not depend on post-matching covariate baseline balance at the district level (it only depends on parallel trends and no alternative treatments), we describe matching results briefly, beginning with the math score sample. Cohort 1 matching examined 234 coarsened strata and matched within four, matching five of six NEE districts to 67 non-NEE districts. Cohort 2 matching used 287 coarsened strata, matched using 16 strata, and matched 19 of 26 NEE districts to 127 non-NEE districts. The mean differences between matched NEE and non-NEE districts across Cohort 1 and 2 districts ranged from -0.03 to 0.03 SD regarding prior-year average student math scores and -$250 to $195 in prior-year PPE.

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

Descriptive District-Level Prior-Year Student Achievement Trends

There is some evidence that pre-intervention achievement trends in *pre-matched* districts that did not adopt NEE throughout the study period are not parallel to trends in districts that implemented NEE; however, graphical analysis suggests that the matching procedure successfully identified comparison districts with trends paralleling NEE district's prior-year student achievement scores. Figure 1 graphs the average district-level achievement scores in NEE, non-NEE, and

matched non-NEE districts. The top-left panel suggests that pre-matched non-NEE and Cohort 1's pre-intervention math score trends are not parallel. While pre-matched non-NEE pre-intervention trends hover around -0.02, Cohort 1's ranges from approximately 0.08 to -0.05. However, the top-right panel shows that Cohort 2's pre-intervention math score trend parallels the pre-matched non-NEE trend. The matching procedure produced prior-year math score trends that parallel NEE trends in each cohort. Moreover, Cohort 1's trend and matched the non-NEE pre-intervention trend are near-equivalent. The bottom-left panel shows that NEE, all non-NEE, and matched non-NEE pre-intervention trends are largely parallel, although NEE district reading scores deviate from the trend four years before NEE implementation. Finally, the bottom-right panel suggests that Cohort 2 pre-intervention trends are parallel and near-equivalent.

Table 2

Descriptive Statistics

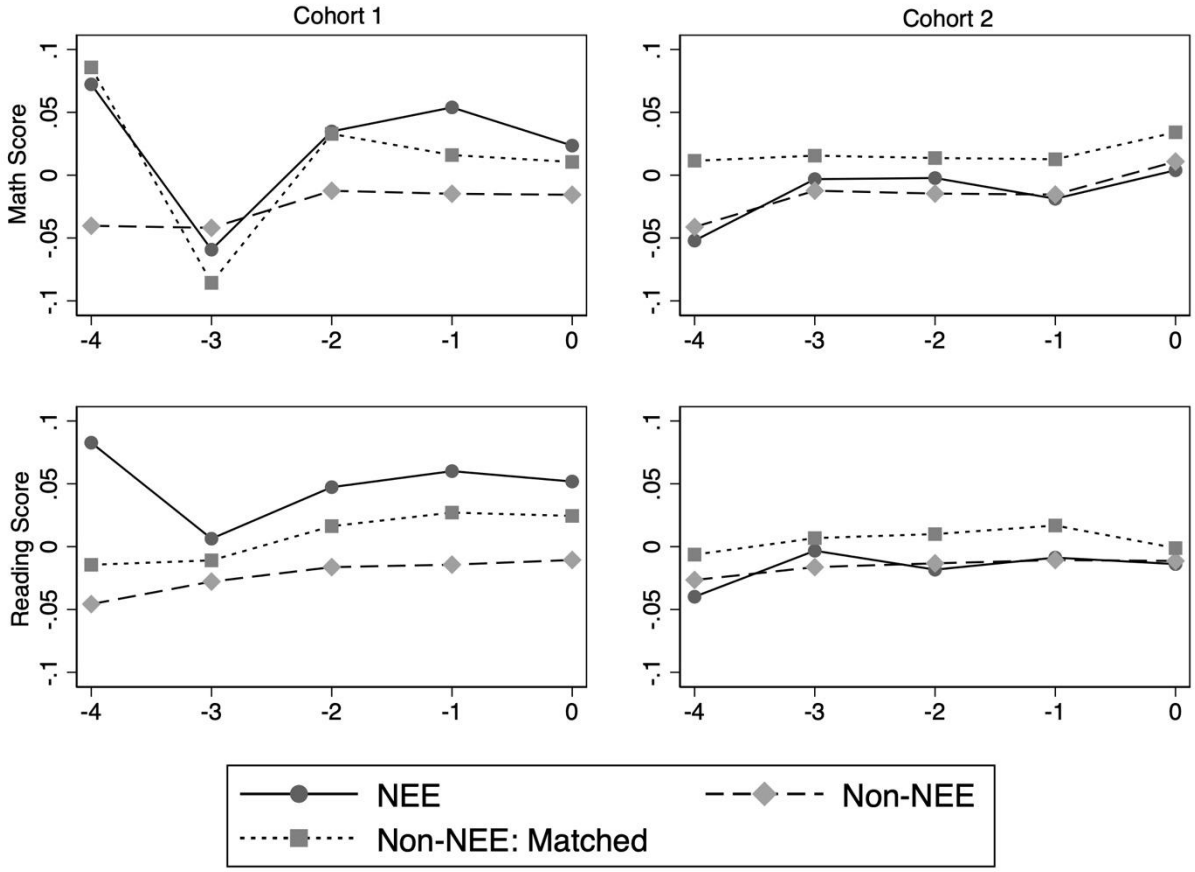
	NEE	Matched and Unmatched Non-NEE
Panel A. Student-Level Characteristics		
Prior-Year Math Score	0.01	0.01
	(0.93)	(0.99)
	[16209]	[470928]
Prior-Year Reading Score	0.02	0.01
	(0.94)	(0.99)
	[16231]	[474234]
Nonwhite	0.11	0.22
	(.)	(.)
	[20535]	[595834]
FRPL	0.54	0.50
	(.)	(.)
	[20535]	[595878]
Panel 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]

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

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

Figure 1

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



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

Post-Matching Generalized DD Results

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

Given the near-zero and statistically insignificant relationships with student achievement scores in both subjects, the rest of the paper focuses on school characteristics correlated with these

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

Table 3

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

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

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

Heterogeneity

Table 4 presents NEE's relationships with student achievement, moderated by the binary indicator regarding school-level average student prior-year achievement and average teacher years of experience. NEE is associated with an increase in student-level math achievement of 0.04 SD in schools where the average student's prior-year achievement score was below the state average (Panel A, Column I). To assess inferential rigor, we test the internal validity of the research design for this subsample of schools. We begin by re-estimating Equation 1 on the subsample of schools with below-average student achievement in math and find a similar relationship (0.06 SD, Panel A, Column II). Notably, the subsample estimate in Column II is insensitive to the use of control variables (Panel A, Column III), suggesting that our research design already accounts for observable differences between NEE and non-NEE schools and unobservable differences that strongly correlate with the observables.

When examining other heterogeneous relationships with math and reading scores, we repeat the moderation, subsample, and subsample with control variables analyses. Table 4, Panel B shows a positive relationship with math scores in schools with below-average teacher years of experience (0.04 SD, Column IV), and this finding holds across the subsample and subsample-with-controls analyses (Panel A Columns V–VI). Similarly, Panel B displays a positive relationship with reading achievement in schools with below-average prior-year student achievement (0.03 SD, Columns I–II); however, we lose precision after adding control variables to the model. Similarly, we find a positive association with reading achievement in schools with below-average teacher years of experience

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

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

Table 4

School Characteristics Moderating NEE's Effects on Achievement Scores

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

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

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 the previously estimated positive relationships with math achievement scores in schools with below-average prior-year student achievement and teacher years of experience and improved reading scores in schools with below-average prior-year student achievement (Appendix E). Pre-intervention differences in achievement across NEE and non-NEE districts are statistically indistinguishable from the score difference in the year before NEE's launch—the omitted category. Appendix E, Panels A and B clearly show stable (parallel) differences in achievement scores between NEE and non-NEE districts in the below-average school subsamples until NEE's introduction, strongly suggesting that post-NEE differences are attributable to NEE. Furthermore, post-NEE event study estimates closely resemble the generalized DD estimates from Table 4. Specifically, among schools with below-average prior-year student achievement, NEE's introduction is associated with increased student-level math of 0.07 SD and reading by 0.03 SD relative to the year before treatment (Appendix E, Panel A). Among schools with below-average teacher years of experience, NEE is associated with math scores that rose by 0.05 SD relative to the year before treatment, though its relationship with reading achievement in these schools is near-zero and statistically insignificant (Appendix E, Panel B).[6]

Between-Cohort Comparisons and Relationships Over Time

Appendix G separates one-year-after-treatment relationships by cohort, estimates one- and two-year-after-treatment relationships for Cohort 1 only, and, for reasons described in the Selection section, compares achievement scores from Cohort 1 against Cohort 2 only.[7] Columns I and IV use the same samples as Appendix E but apply a version of Equation 1 where we interact a cohort identifier with the treatment indicator. Appendix G, Panel A Column I shows positive one-year-after changes in math achievement scores among schools with below-average student achievement across both cohorts; however, only Cohort 2's change is statistically significant. Among schools with below-average teacher years of experience, scores declined in NEE Cohort 1 and rose in Cohort 2, but only the latter is statistically significant (Appendix G, Panel B Column I). Relationships with reading are somewhat similar. Among schools with below-average prior-year achievement scores, NEE's introduction is associated with 0.03 SD higher reading achievement in both cohorts, but this time, each estimate is statistically significant (Appendix G, Panel A Column IV). Like the relationships with math scores in schools with below-average teacher years of experience, there is a drop in reading achievement in Cohort 1 and a rise in Cohort 2; however, neither is statistically significant (Appendix G, Panel B Column IV). Notably, the evidence in Columns I and IV is inconsistent with positive selection on perceived superintendent effectiveness (see Selection section).

Next, we turn to one- and two-year-after relationships for Cohort 1; these analyses restrict the samples from Appendix E to only Cohort 1 and its matches and apply a version of Equation 1 in which we interact treatment with a variable indicating whether the data were collected one or two years after treatment (see Appendix G Columns II and V). Math scores in NEE Cohort 1 schools with below-average prior-year student achievement, compared to below-average non- or not-yet-NEE schools, suggest that NEE's associations with math scores may fade over time (Appendix G,

[6] We also apply event studies to each content-specific sample from Table 3, which includes below and above-average schools (see Appendix Table F1). Again, we detect no post-treatment relationships, and the data corroborate the parallel trends assumption.

[7] See Appendix Table F2 for corresponding analyses using the full sample from which results in Table 3 are based.

Panel A Colum II). However, the opposite occurs among schools with below-average teacher years of experience (Panel B Column II): one year after NEE's introduction, math scores decline, though the change is statistically insignificant, and scores rise by a statistically significant 0.06 SD two years after NEE's introduction. While we know the precise type of human capital deficiency in schools with below-average teacher experience (i.e., on-the-job experience), schools with below-average prior-year student achievement may suffer from multiple human capital deficiencies. At face value, the results in Column II may suggest that NEE can immediately address the various levels of human capital needs in low-performing schools. At the same time, high concentrations of teacher inexperience may be less tractable, and therefore, schools with such teachers need prolonged NEE exposure before realizing improvement. Results in Column V show that reading scores change in the same direction as math scores, though no estimates in either panel are statistically significant.

Finally, we examine results from another indirect test of positive selection by comparing NEE Cohort 1 schools to not-yet-treated NEE Cohort 2 schools only, leveraging variation in treatment timing (Appendix G, Columns III and VI). Recall that substantive positive selection on perceived superintendent effectiveness would cause achievement score differences between Cohort 1 and non-NEE schools to exceed those between Cohort 1 and 2 (see Selection section). Among schools with below-average prior-year student achievement, the difference between Cohort 1 and non-NEE and not-yet-NEE math achievement is 0.07 SD (Table 4 Panel A Column III), while the difference between Cohort 1 and 2 is 0.09 SD (Online Appendix G Panel A Column III). The corresponding differences in math scores among schools with below-average teacher years of experience are 0.04 SD (Table 4 Panel A Column VI) and 0.07 SD (Appendix G Panel B Column III). These patterns are inconsistent with the assumption regarding positive selection on perceived superintendent effectiveness, affirming the research design's internal validity concerning math scores. However, the evidence regarding reading scores is less compelling; the estimates in Table 4 Panel B Columns III and VI are small and statistically insignificant, and those in Appendix G Panels A and B Column VI are smaller and insignificant.

Sensitivity Analyses

Formal sensitivity tests suggest that our inferences are insensitive to plausible OVs, bolstering our confidence in the internal validity of our research design. We examine OVs with up to 30% of the explanatory power of student-by-year prior-year achievement since an OV mimicking this relationship explains 95.4% of the residual variation in math scores in schools with below-average achievement that is not explained by our model (Appendix H Panel A Column II). Appendix H Panel A presents sensitivity analyses regarding the effects of our hypothetical OVs on math and reading scores for those schools with below-average prior-year student achievement. While none of the benchmarked OVs explain a substantial amount of residual treatment variation, student-by-year prior-year scores explain substantial amounts of residual outcome variation; however, an OV resembling this powerful predictor of residual outcome variation has virtually no effect on our inferences (Appendix H Panel A Columns III and VI). We observe similar patterns among schools with below-average teacher years of experience (Panel B Column III) and note that the effects of our OVs on reading scores, which have not resulted in significant relationships in any model, remain statistically insignificant.

Placebo Tests and Relationships with Covariates

Placebo tests affirm the internal validity of the research design for estimating math score relationships but present less compelling evidence for the design estimating reading score relationships among schools with below-average prior-year achievement (Appendix I Table I1). Panels A1 and A2 detect no placebo relationships with math scores among either subsample of

schools. However, results in Panel B1 asperse the internal validity of the research design for reading scores when applied to schools with below-average prior-year student achievement, though we detect no placebo relationships on reading scores among schools with below-average teacher years of experience (Panel B2). Given the lack of detected relationships with reading scores and evidence aspersing the internal validity of our research design when applied to reading subsamples, the remainder of the paper focuses on math scores. Appendix Table I2 displays placebo test results using the full samples from Table 3 and detects no placebo effects.

As a developmentally focused teacher evaluation system eschewing evaluation for accountability, NEE is not designed to alter the compositions of students or teachers in its schools or districts. Consequently, we should not detect any relationships with variables describing these compositions; if we find compositional changes, it could suggest the presence of alternative treatments (that aim to change compositions). Appendix Tables I3 and I4 present no evidence suggesting that NEE introduced compositional changes among the math subsamples comprised of schools with below-average prior-year student achievement or teacher years of experience, once again authenticating the internal validity of our research design for math scores. However, we detect some compositional changes among schools with below-average prior-year student achievement in reading (Appendix Table I3). While the number of changes we detect across Tables I3 and I4 could arise from Type I error, we conclude that the evidence collectively undermines the internal validity of our research design for detecting NEE's relationships with reading scores. We also note that we detect no compositional changes among the full math and reading samples used in Table 3 (see Appendix Table I5).

Conclusion

There is insufficient evidence to reach defensible conclusions regarding the average effects of teacher evaluation on student achievement, less evidence about the conditions in which evaluation works and its cost-effectiveness, and no rigorous work focused on rural settings, leaving most local policymakers in the dark. We addressed these gaps by examining a rurally focused, fee-based teacher evaluation system, the Network for Educator Effectiveness (NEE). Furthermore, despite the availability of alternative teacher evaluation systems without fees, NEE remains popular amongst districts, underscoring its relevance.

We first compared NEE's fees to other teacher evaluation systems and found that NEE's $3 per student per year membership fee is lower, possibly substantially lower than districts would otherwise pay to implement a comparable system. NEE's cost advantage is particularly significant given the resource constraints facing rural districts, which often lack the infrastructure, specialized personnel, and economies of scale that enable urban districts to implement comprehensive evaluation systems (Jacques et al., 2000). Like prior work concerning teacher evaluation costs (Chambers et al., 2013), we do not assert that NEE's fee captures all costs required to design, implement, and maintain a teacher evaluation system. Nevertheless, NEE's fee arguably represents the cost of interest to districts. Moreover, linking the relationships of evaluation to its fees, despite their limited information about indirect costs, is an improvement over prior work that only reports effects.

However, on average, NEE does not affect rural student math or reading achievement. All point estimates from our full samples (Table 3) are statistically insignificant, near-zero, and negligible according to work regarding effect size interpretation (Jacob et al., 2019; Kraft, 2020). Our conclusion regarding average relationships is consistent with prior work from urban and national settings, which finds no math or reading effects or subject-specific effects (Bleiberg et al., 2024; Steinberg & Sartain, 2015; Taylor & Tyler, 2012). While the absence of average relationships

suggests that rural policymakers might not unconditionally adopt NEE or systems like it, this does not mean that NEE never improves rural student achievement. Indeed, this is why we pushed beyond average relationships.

NEE may improve reading achievement in rural schools where the average student's prior-year achievement score is below the state average or the average teacher's years of experience are below the state average; the evidence repeatedly suggests that NEE improves math achievement in these schools. Although reading achievement increased in NEE schools with below-average prior-year achievement or teacher years of experience, we could not rule out plausible confounders. However, the data never suggested threats to the internal validity of estimates regarding math scores. Among rural schools with below-average prior-year student achievement, NEE is associated with modest but statistically significant improvements in math scores of approximately 0.07 SD or 2.3 months of learning and higher math scores of about 0.04 SD or 1.3 months of student learning in rural schools where the typical teacher had below-average years of experience.[8] These improvements are particularly noteworthy given the observer capacity challenges in rural schools, where administrators often juggle multiple roles and maintain closer personal relationships with teachers - dynamics that can complicate objective evaluation (Eranil, 2023; Preston et al., 2018). Importantly, these one-year-after relationships are plausible and representative of estimates from prior work in urban settings, which also focus on one-year-after relationships (Steinberg & Sartain, 2015; Taylor & Tyler, 2012). Notably, NEE's relationships-to-expenditure ratios in these below-average rural schools range from 0.013 SD to 0.023 SD per dollar spent per student. To place these ratios in context, Harter (1999) reports that increasing teacher salary supplements by $1 per teacher (in 2012 dollars) is associated with an increase in student math achievement scores of 0.0006 SD, and Wenglinsky (1997) finds that increasing PPE assigned to the broad category of "instructional expenditures" by one 2012 dollar is associated with a rise of 0.000003 SD in mathematics.

Analyses regarding NEE's relationships with math scores in the below-average rural schools examined reveal context-dependent variability. In rural schools with below-average prior-year achievement, NEE is associated with higher math achievement one year after treatment, implying that NEE may be an enticing and cost-effective intervention for below-average schools facing immediate pressure to improve math scores. Simultaneously, positive relationships in schools with high concentrations of inexperienced teachers took two years. Schools with large concentrations of inexperienced teachers must impart substantial professional knowledge to their staff while contending with geographic isolation, limited opportunities for peer collaboration, and barriers to accessing specialized training (Erickson et al., 2012; Hansen-Thomas & Grosso Richins, 2015; Nugent et al., 2016; Quintana et al., 2000). We speculate that NEE—which introduces new teacher performance expectations and measures and substantially increases the number of performance feedback episodes received—may initially overwhelm an inexperienced staff already acquiring substantial professional knowledge. Teachers may need additional time to incorporate this knowledge into practice. While we believe NEE can help schools with large concentrations of inexperienced teachers improve math achievement, leaders in these settings should not expect immediate benefits.

Our study examined school settings like those in Steinberg and Sartain's (2015) analysis of Chicago teacher evaluation (2015). While Steinberg & Sartain found that introducing a reformed teacher evaluation system may exacerbate inequality, we found that it may benefit schools that need improvement. We conjecture that these diametric results stem from the value-add of a new

[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.

developmentally oriented intervention (e.g., Chicago's EITP and Missouri's NEE) in urban versus rural settings. If we assume that Chicago already targeted robust professional development to the schools needing the most improvement (i.e., those with below-average student achievement and less experienced teachers), adopting another developmentally focused intervention (EITP) may push these schools toward diminishing marginal returns. However, rural schools, which often receive less robust and frequent professional development than urban schools (Skyhar, 2020), may still be at a point where a new developmentally focused intervention (NEE) results in rising marginal returns. Future work might test the validity of this conjecture.

Limitations

This study may be limited in several ways. First, our findings may not generalize to other settings. NEE receives input from faculty experts in teacher evaluation and some university-based financial support, which may affect NEE's deliverables; such expertise and financial resources may not be available in other systems. NEE was also designed for rural schools, and all data came from these schools; as we argued above, the dynamics of rural teacher labor markets and smaller central office infrastructures alone may affect our data in ways that would not apply in urban settings. Additionally, the design of teacher evaluation systems varies substantially across systems (Bleiberg et al., 2024); systems based on different design elements than NEE may not realize similar results. Second, NEE's relationships with math (and potentially reading) scores in the below-average school settings examined may change over longer periods. Future work with longer panels might explore these relationships in rural settings.

Third, despite the robustness of the math estimates to the threats examined, we cannot rule out all possible confounders (no quasi-experiment can). The voluntary adoption of NEE raises concerns about selection bias that our research design may not fully address. Districts that chose NEE may have systematically differed from non-adopters in leadership capacity, openness to innovation, or prior instructional quality. Our sensitivity analyses provide critical readers with information about what must be true about inference-undoing omitted variables in our research design, allowing them to make evidence-based decisions regarding internal validity and whether inference-undoing threats are likely in specific contexts. For example, an unobserved confounder must explain residual variation in both treatment assignment and student outcomes equivalent to 30% of the explanatory power of prior-year achievement scores to undo our inferences. Given that prior-year achievement captures the cumulative effect of most factors affecting student learning, we argue that such a confounder operating independently of all our measured covariates is unlikely, although we cannot definitively rule it out.

Fourth, we examine only student achievement outcomes while our theoretical framework emphasizes that NEE likely operates via developmental mechanisms including improved teaching quality, enhanced professional development, and better teacher retention. Future research should examine such mechanisms to provide a more complete understanding of how and why NEE and other teacher evaluation systems affect student achievement, especially because teacher evaluation reforms are often justified on the grounds that they improve educator outcomes. While our student achievement findings are an important contribution to understanding NEE's and, more broadly, teacher evaluation's effects, they provide an incomplete picture of the full range of outcomes that evaluation systems are designed to influence. Indeed, we assume that NEE's users believe it affects important unexamined outcomes positively; otherwise, we cannot fathom why districts would choose to join the fee-based NEE system. Finally, as discussed, we report relationships-to-expenditure ratios, falling short of the ideal cost-effectiveness ratios.

Implications

Our work affords targeted policy implications, which we offer while urging caution befitting a single study. We do not advise rural districts to adopt systems like NEE unconditionally; instead, it may be better to consider these systems as interventions to help rural schools with below-average prior-year student achievement and teacher years of experience improve math (and potentially) reading scores. We also encourage education agencies serving rural schools to embed professional growth opportunities within the evaluation system (through feedback, goal-setting, and evaluation-informed professional support), which may be particularly valuable for rural schools struggling to access external PD.

Knowing NEE's fee can also help rurally minded policymakers choose the right intervention for improving achievement scores. The fee districts pay to NEE is associated with math achievement score improvements between 0.04 and 0.07 SD, suggesting that NEE is incredibly cost-effective in these settings based on short-term outcomes. Indeed, policymakers seeking immediate improvements in math scores in these contexts may find NEE a cost-effective short-term option, although its longer-term effectiveness remains unknown. State policymakers might consider supporting the creation of university-based models like NEE as a potentially cost-effective approach for short-term rural achievement gains. Rather than each rural district trying to build capacity independently, shared infrastructure through a central hub shows initial promise as a cost-effective delivery model. However, these cost-effectiveness assessments are based on one-year post-implementation effects, and whether NEE's benefits persist, fade, or strengthen over time remains an empirical question.

Acknowledgements

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

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/

Bacher-Hicks, A., & Koedel, C. (2023). Estimation and interpretation of teacher value added in research applications. In Hanushek, E. A., Machin, S. & Woessmann, L. (Eds.), *Handbook of the economics of education* (Vol. 6, pp. 93-134). Elsevier.

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

Bleiberg, J., Brunner, E., Harbatkin, E., Kraft, M. A., & Springer, M. G. (in press). The effect of teacher evaluation on achievement and attainment: Evidence from statewide reforms. *Journal of Political Economy Microeconomics*. https://doi.org/10.1086/732837

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

Chambers, J., Brodziak de los Reyes, I., & O'Neil, C. (2013). *How much are districts spending to implement teacher evaluation systems? Case studies of Hillsborough County Public Schools, Memphis City Schools, and Pittsburgh Public Schools* (Working Paper No. WR-989-BMGF). 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?* (No. REL 2017-190; Making Connections, pp. 1–29). REL Central.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). The long term impacts of teachers: Teacher value added and student outcomes in adulthood. *American Economic Review*, *104*(9), 2633–2679.

Choi, E., & Johnson, D. A. (2022). Common antecedent strategies within organizational behavior management: The use of goal setting, task clarification, and job aids. *Journal of Organizational Behavior Management*, *42*(1), 75–95. https://doi.org/10.1080/01608061.2021.1967834

Cinelli, C., & Hazlett, C. (2020). Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *82*(1), 39–67. https://doi.org/10.1111/rssb.12348

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2006). Teacher-student matching and the assessment of teacher effectiveness. *The Journal of Human Resources*, *41*(4), 778–820.

Cowan, J., Goldhaber, D., & Theobald, R. (2022). Performance evaluations as a measure of teacher effectiveness when implementation differs: Accounting for variation across classrooms, schools, and districts. *Journal of Research on Educational Effectiveness*, *15*(3), 510–531. https://doi.org/10.1080/19345747.2021.2018747

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.

Eranil, A. K. (2023). Dwelling upon the backstage of the success of rural schools: A systematic review analysis. *International Journal of Educational Research Review*, *8*(4), 832–852. https://doi.org/10.24331/ijere.1324029

Erickson, A. S. G., Noonan, P. M., & Mccall, Z. (2012). Effectiveness of online professional development for rural special educators. *Rural Special Education Quarterly*, *31*(1), 22–32. https://doi.org/10.1177/875687051203100104

Feng, L. (2010). Hire today, gone tomorrow: New teacher classroom assignments and teacher mobility. *Education Finance and Policy*, *5*(3), 278–316. https://doi.org/10.1162/EDFP_a_00002

Frahm, M., & Cianca, M. (2021). Will they stay or will they go? Leadership behaviors that increase teacher retention in rural schools. *The Rural Educator*, *42*(3), 1–13. https://doi.org/10.35608/ruraled.v42i3.1151

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

Gormley, T. A., & Matsa, D. A. (2011). Growing out of trouble? Corporate responses to liability risk. *Review of Financial Studies*, *24*(8), 2781–2821. https://doi.org/10.1093/rfs/hhr011

Hansen-Thomas, H., & Grosso Richins, L. (2015). ESL mentoring for secondary rural educators: Math and science teachers become second language specialists through collaboration. *TESOL Journal*, *6*(4), 766–776. https://doi.org/10.1002/tesj.221

Harter, E. A. (1999). How educational expenditures relate to student achievement: Insights from Texas elementary schools. *Journal of Education Finance*, *24*(3), 281–302.

Hunter, S. B. (2020). The unintended effects of policy-assigned teacher observations: Examining the validity of observation scores. *AERA Open*, *6*(2). https://doi.org/10.1177/2332858420929276

Hunter, S. B., & Ege, A. (2021). Linking student outcomes to school administrator discretion in the implementation of teacher observations. *Educational Administration Quarterly*, *57*(4), 607–640. https://doi.org/10.1177/0013161X211003134

Hunter, S. B., & Kho, A. (2025). The effects of teacher evaluation policy on student achievement and teacher turnover: Leveraging teacher accountability and teacher development. *Journal of Education Human Resources*, *43*(3), 582–633. https://doi.org/10.3138/jehr-2023-0040

Hunter, S. B., & Springer, M. G. (2022). Performance feedback, human capital, and teacher performance: A mixed-methods analysis. *Educational Evaluation and Policy Analysis*, *44*(3), 380–403. https://doi.org/10.3102/01623737211062913

Hunter, S. B., & Steinberg, M. P. (2024). *The valence of teacher performance feedback and its consequences: Examining a critical mechanism of reformed teacher evaluation system.* [EdWorkingPaper 22-676]. Annenberg Institute, Brown University. https://doi.org/10.26300/97k9-br18

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

Ingersoll, R. M., & Tran, H. (2023). Teacher shortages and turnover in rural schools in the US: An organizational analysis. *Educational Administration Quarterly*, *59*(2), 396–431. https://doi.org/10.1177/0013161X231159922

Jackson, C. K. (2018). What do test scores miss? The importance of teacher effects on non–test score outcomes. *Journal of Political Economy*, *126*(5), 2072–2107. https://doi.org/10.1086/699018

Jacob, R. T., Doolittle, F., Kemple, J., & Somers, M.-A. (2019). A framework for learning from null results. *Educational Researcher*, *48*(9), 580–589. https://doi.org/10.3102/0013189X19891955

Jacques, C., Brorsen, B. W., & Richter, F. G. (2000). Consolidating rural school districts: Potential savings and effects on student achievement. *Journal of Agricultural and Applied Economics*, *32*(3).

Kalogrides, D., & Loeb, S. (2013). Different teachers, different peers: The magnitude of student sorting within schools. *Educational Researcher*, *42*(6), 304–316. https://doi.org/10.3102/0013189X13495087

Kimball, S. M., & Milanowski, A. (2009). Examining teacher evaluation validity and leadership decision making within a standards-based evaluation system. *Educational Administration Quarterly*, *45*(1). https://doi.org/10.1177/0013161X08327549

Kraft, M. A. (2019). Teacher effects on complex cognitive skills and social-emotional competencies. *Journal of Human Resources*, *54*(1), 1–36. https://doi.org/10.3368/jhr.54.1.0916.8265R3

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

Kraft, M. A., & Gilmour, A. F. (2016). 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. (2017). Revisiting the widget effect: Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Researcher*, *46*(5), 234–249. https://doi.org/10.3102/0013189X17718797

Liebowitz, D. D., Porter, L., & Bragg, D. (2022). The effects of higher-stakes teacher evaluation on office disciplinary referrals. *Journal of Research on Educational Effectiveness*, *15*(3), 475–509. https://doi.org/10.1080/19345747.2021.2015496

Liu, J., & Loeb, S. (2021). Engaging teachers: Measuring the impact of teachers on student attendance in secondary school. *Journal of Human Resources*, *56*(2), 343–379. https://doi.org/10.3368/jhr.56.2.1216-8430R3

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

Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives*. Sage Publications.

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

Nguyen, T. D., Pham, L. D., Crouch, M., & Springer, M. G. (2020). The correlates of teacher turnover: An updated and expanded meta-analysis of the literature. *Educational Research Review*, *31*, 100355. https://doi.org/10.1016/j.edurev.2020.100355

Nugent, G. C., Chumney, F. L., Ihlo, T., Shapiro, E. S., Guard, K., Koziol, N., & Bovaird, J. (2016). Investigating rural teachers' professional development, instructional knowledge, and classroom practice. *Journal of Research in Rural Education*, *31*(3), 1–16.

Papay, J. P. (2012). Refocusing the debate: Assessing the purposes and tools of teacher evaluation. *Harvard Educational Review*, *82*(1), 123–141. https://doi.org/10.17763/haer.82.1.v40p0833345w6384

Phipps, A. (2022). *Does monitoring change teacher pedagogy and student outcomes?* [EdWorkingPaper, 22–510]. Annenberg Institute at Brown University. https://doi.org/10.26300/7021-1x97

Phipps, A., & Wiseman, E. A. (2021). Enacting the rubric: Teacher improvements in windows of high-stakes observation. *Education Finance and Policy*, *16*(2), 283–312. https://doi.org/10.1162/edfp_a_00295

Preston, J. P., Jakubiec, B. A. E., & Kooymans, R. (2018). Common challenges faced by rural principals: A review of the literature. *The Rural Educator*, *35*(1). https://doi.org/10.35608/ruraled.v35i1.355

Quintana, C., Krajcik, J., & Soloway, E. (2000). Exploring a structured definition for learner-centered design: A definition for learner-centered design. In B. J. Fishman & S. F. O'Connor-Divelbiss (Eds.), *International Conference of the Learning Sciences: Facing the Challenges of Complex Real-world Settings* (pp. 256–263). Psychology Press.

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. (2020). Understanding tenure reform: An examination of sense-making among school administrators and teachers. *Teachers College Record*, *122*(11), 42. https://doi.org/10.1177/016146812012201112

Rodriguez, L. A., Swain, W. A., & Springer, M. G. (2020). Sorting through performance evaluations: The influence of performance evaluation reform on teacher attrition and mobility. *American Educational Research Journal*, 000283122091098. https://doi.org/10.3102/0002831220910989

Rosenbaum, P. R., & Rubin, D. B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society Series B*, *45*(2), 212–218. https://doi.org/10.1111/j.2517-6161.1983.tb01242.x

Schwartz, N., Kang, H., Loeb, S., Grissom, J., Bartanen, B., Cheatham, J., Chi, O., Donaldson, M., Lemos, R. F., Mellon, G., Moffitt, S., Nurshatayeva, A., Owens, J., Pinker, E., White, R., & Zimmerman, S. (2023). *Studying the superintendency: A call for research.* Annenberg Institute at

Brown University. https://annenberg.brown.edu/sites/default/files/Studying%20the%20Superintendency%20-%20Call%20for%20Research.pdf

Skyhar, C. (2020). Thinking outside the box: Providing effective professional development for rural teachers. *Theory & Practice in Rural Education*, *10*(1), 42–72. https://doi.org/10.3776/tpre.2020.v10n1p42-72

Song, M., Wayne, A. J., Garet, M. S., Brown, S., & Rickles, J. (2021). Impact of providing teachers and principals with performance feedback on their practice and student achievement: Evidence from a large-scale randomized experiment. *Journal of Research on Educational Effectiveness*, 1–26. https://doi.org/10.1080/19345747.2020.1868030

Stecher, B. M., Garet, M. S., Hamilton, L. S., Steiner, E. D., Robyn, A., Poirier, J., Holtzman, D., Fulbeck, E. S., Chambers, J., & Brodziak de los Reyes, I. (2016). *Improving teaching effectiveness* (No. 9780833092212). RAND. https://www.rand.org/content/dam/rand/pubs/research_reports/RR1200/RR1295/RAND_RR1295.pdf

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

Tran, H., Hardie, S., Gause, S., Moyi, P., & Ylimaki, R. (2020). Leveraging the perspectives of rural educators to develop realistic job previews for rural teacher recruitment and retention. *The Rural Educator*, *41*(2), 31–46. https://doi.org/10.35608/ruraled.v41i2.866

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect*. The New Teacher Project (TNTP). http://tntp.org/assets/documents/TheWidgetEffect_2nd_ed.pdf

Wenglinsky, H. (1997). School district expenditures, school resources and student achievement: Modeling the production function. In W. J. Jr. Fowler (Ed.), *Developments in school finance 1997* (p. 196). National Center for Education Statistics.

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

About the Authors

Seth B. Hunter
George Mason University
shunte@gmu.edu
https://orcid.org/0000-0002-3051-872X
Seth B. Hunter is an associate professor of educational leadership and policy at George Mason University. He applies psychometrics, econometrics, computational, and mixed methods to examine educator and organizational effectiveness, human–machine partnerships, and research use by practitioners and policymakers.

Katherine M. Bowser
National Council on Teacher Quality
https://orcid.org/0009-0004-4771-2272
Katherine M. Bowser is a researcher at the National Council on Teacher Quality. She applies quantitative methods to examine teacher effectiveness and educator labor markets, particularly as influenced by K–12 district policy.

education policy analysis archives

Volume 34 Number 42 May 5, 2026 ISSN 1068-2341