ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 21-496

The Effect of Teacher Evaluation on Achievement and Attainment: Evidence from Statewide Reforms

Joshua Bleiberg
Brown University

Eric Brunner
University of Connecticut

Erica Harbatkin
Michigan State University

Matthew A. Kraft
Brown University

Matthew G. Springer
University of North Carolina at Chapel Hill

Starting in 2009, the U.S. public education system undertook a massive effort to institute new high-stakes teacher evaluation systems. We examine the effects of these reforms on student achievement and attainment at a national scale by exploiting the staggered timing of implementation across states. We find precisely estimated null effects, on average, that rule out impacts as small as 1.5 percent of a standard deviation for achievement and 1 percentage point for high school graduation and college enrollment. We also find little evidence of heterogeneous effects across an index measuring system design rigor, specific design features, and district characteristics.

VERSION: December 2021

Suggested citation: Bleiberg, Joshua, Eric Brunner, Erica Harbatkin, Matthew A. Kraft, and Matthew Springer. (2021). The Effect of Teacher Evaluation on Achievement and Attainment: Evidence from Statewide Reforms. (EdWorkingPaper: 21-496). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/b1ak-r251

The Effect of Teacher Evaluation on Achievement and Attainment: Evidence from Statewide Reforms

Joshua Bleiberg[1], Eric Brunner[2], Erica Harbatkin[3], Matthew A. Kraft[1], and Matthew G. Springer[4]

Abstract

Starting in 2009, the U.S. public education system undertook a massive effort to institute new high-stakes teacher evaluation systems. We examine the effects of these reforms on student achievement and attainment at a national scale by exploiting the staggered timing of implementation across states. We find precisely estimated null effects, on average, that rule out impacts as small as 1.5 percent of a standard deviation for achievement and 1 percentage point for high school graduation and college enrollment. We also find little evidence of heterogeneous effects across an index measuring system design rigor, specific design features, and district characteristics.

JEL Codes: I20, I28, J21

Keywords: Teacher evaluation, Student achievement, Education

Corresponding author, Joshua Bleiberg, can be reached at joshua_bleiberg@brown.edu. Authors are listed in alphabetical order. The Spencer Foundation [Award#201700052] and the Institute for Education Sciences [Award # R305A170053] provided generous support to Matthew Kraft for this work. We are grateful for the feedback from Melissa Lyon, Danielle Edwards, Grace Falken, Alvin Christian, Alex Bolves, the participants in the Brown Half-Baked Research Series, the Annual Northeast Economics of Education Workshop, and the Society for Research on Educational Effectiveness annual conference.

1 *Brown University*, 2 *University of Connecticut, 3 Michigan State University, and 4 University of North Carolina at Chapel Hill.*

The returns to improved performance evaluation systems have long been of interest to economists and employers. Evaluation systems have the potential to better align worker's effort with organizational goals as well as to inform employee skill development (Oyer and Schaefer 2011; Prendergast 1999; Gibbons 1998). We study the effects of performance evaluation in the K-12 public education system, which with more than 3.5 million teacher employees is one of the largest economic sectors in the U.S. Research demonstrates that teachers have large effects on a range of student outcomes, but that teacher effectiveness varies considerably (Chetty, Friedman, and Rockoff 2014; Jackson 2018; Kraft 2019; Petek and Pope 2016). Understanding the impacts of more rigorous and regular performance reviews for public school teachers is particularly important given the sizable potential gains from improving teacher productivity.

Between 2009 and 2017, 44 states and Washington, D.C. implemented major reforms to their teacher evaluation systems. Prior to the reforms, teacher evaluation was largely a perfunctory exercise that resulted in nearly all teachers receiving satisfactory ratings (Weisberg et al. 2009). Strong incentives by the federal government helped spur the widespread reforms. The $4.35 billion federal Race to the Top (RTTT) grant competition incentivized states to reform evaluation systems by regularly evaluating teachers based on multiple measures (including student academic growth) and using performance ratings to inform personnel decisions.

The new evaluation systems were highly controversial, leading to protests and lawsuits challenging their legitimacy in several states (Government Accountability Office 2015; McGuinn 2012; Sawchuk 2015). Proponents argued that reforming teacher evaluation systems would allow districts to attract and retain more effective teachers by closely linking personnel decisions and compensation to rigorous, multi-measure teacher evaluation ratings (Hanushek 2009). Opponents argued that the new high-stakes evaluation systems were based

on invalid and unreliable metrics that would disincentivize cooperation and make the profession less attractive to prospective teachers (Murphy, Hallinger, and Heck 2013). The reforms were also financially costly, with a conservative estimate of $15-20 billion dollars from 2012 to 2018 (Chambers, Brodziak de los Reyes, and O'Neil 2013).

In this paper, we examine how new teacher evaluation systems taken to scale nationally affected student achievement and educational attainment. Existing evidence on the effects of evaluation reforms in a narrow set of districts is mixed (e.g. Dee and Wyckoff 2015; Taylor and Tyler 2012; Stecher et al. 2018). Further, a recent study using a similar identification strategy as ours found that state-level evaluation reforms raised the quality of new teachers but also decreased their job satisfaction and lowered the overall supply of newly licensed teacher candidets (Kraft et al. 2020). Thus, the net effect of evaluation reforms on student outcomes remains unclear.

We leverage variation in the timing of adoption of new teacher evaluation systems across states to identify the causal effects of these reforms in an event study and difference-in-differences (DiD) framework. We further explore potential heterogeneity in these effects given the substantial variation in the evaluation metrics and design features adopted by states. Our primary analyses combine data on the timing of state adoption of teacher evaluation reforms with comprehensive district-level student achievement data from 2009 to 2018 on standardized math and English Language Arts (ELA) exams from the Stanford Education Data Archive (SEDA). We augment this achievement data with data from the American Community Survey (ACS) Public Use Microdata Sample (PUMS) to examine the impact of teacher evaluation reforms on longer-run student attainment outcomes, namely high school graduation and college enrollment.

We then examine the robustness of our primary specifications using several newly developed two-way fixed effects (TWFE) estimators to test the robustness of our results to heterogeneous treatment effects (Callaway and Sant'Anna 2020; Goodman-Bacon 2021; Sun and Abraham 2020). As a further robustness check, we show that our results are essentially unchanged when controlling for the adoption of related teacher accountability reforms and a wide range of other time-varying education reform efforts.

We find that, on average, state teacher evaluation reforms had no discernable effect on student achievement in math or ELA. Estimates from event study models are small in magnitude and statistically insignificant up to five years post-reform. Further, estimates from DiD specifications produce precisely estimated null effects on achievement; we can rule out positive effects of the reforms as small as 0.015 standard deviations in math and 0.009 standard deviations in ELA. We also find no evidence that teacher evaluation reforms impacted high school graduation or college enrollment rates and can rule out positive effects as small as 1 percentage point for both attainment measures.

We further examine whether our average estimates mask important heterogeneity in treatment effects based on variation in system design across states. We use data on 10 teacher evaluation policy components commonly identified as key features of effective systems (Bleiberg and Harbatkin 2020; Doherty and Jacobs 2015; Howell and Magazinnik 2017).[1] We construct an index of evaluation system design rigor based on the number of features that a state required districts to enact. We then use that index to examine whether there is heterogeneity associated with the rigor of the evaluation system design.

For math scores, we find little evidence that the effect of teacher evaluation reform varied with design rigor. For ELA scores, we find some

[1] See Appendix Table B1 for a full list of the components and their sources.

evidence that states that adopted the fewest evaluation components experienced a slight decline in student achievement post-reform. In contrast, those that adopted the most policy components experienced little change in student achievement. We complement our index approach by grouping system design elements into three broad categories motivated by the primary mechanisms through which proponents argued evaluation would benefit students. We find no evidence of heterogeneity based on these broad categories of evaluation system design. Finally, we test for heterogeneous treatment effects across student body characteristics and find little evidence that teacher evaluation reforms impacted student achievement or attainment for any subgroup.

Our paper makes four primary contributions to the literature. First, we add new empirical evidence to the large performance management literature in economics (Heinrich, Meyer, and Whitten 2010; Heinrich and Marschke 2010; Bloom and Van Reenen 2011; G. Baker 1992; Cappelli and Conyon 2018; Gibbons 1998; Bloom and Van Reenen 2007). Second, our nationally representative study provides the broadest and most generalizable evidence on the efficacy of teacher evaluation reforms in the U.S. Several studies of evaluation systems implemented in individual districts, such as Washington, D.C., and Chicago Public Schools, provide evidence that teacher evaluation reforms have the potential to improve student achievement, but the findings from those studies may lack generalizability (Adnot et al. 2016; Dee, James, and Wyckoff 2021; Dee and Wyckoff 2015; Dotter, Chaplin, and Bartlett 2021; Sartain and Steinberg 2016; Steinberg and Sartain 2015; Taylor and Tyler 2012).[2] We illustrate how our results are consistent with this prior evidence by confirming that a small group of evaluation systems identified ex-post as exemplary did appear to raise student

[2] At the same time, other studies that focus on individual districts or states have found mixed or null effects of teacher evaluation reforms (Anderson, Cowen, and Strunk 2019; Cullen, Koedel, and Parsons 2021; Stecher et al. 2018).

achievement. Third, we provide the first evidence on how teacher evaluation reforms affected students' longer-term outcomes, which speaks to the multiple pathways through which improved teacher effectiveness might benefit students. Finally, we contribute to the cross-disciplinary literature on the efficacy of scaling up promising programs: where several studies in the education literature point to the pitfalls that prior reforms have faced when taken to scale across the decentralized U.S. public education system (Honig 2006; Coburn 2003; Manna 2010; Pressman and Wildavsky 1984; Zhou et al. 2021; Gupta et al. 2021).

The paper proceeds as follows. Section II describes the history and background of teacher evaluation reforms and reviews the related literature. Section III describes the data we assemble to examine the impact of teacher evaluation reforms on student achievement and educational attainment. Section IV outlines our empirical framework for isolating the causal effects of evaluation reforms on our outcomes of interest. We present our main findings in Section V and conclude in Section VI with a discussion of the implications of our results for policy and practice.

II. Background

The widespread adoption of teacher evaluation reforms marked a shift from evaluation systems that relied primarily on teacher observation and typically had little, if any, connection with teacher compensation or employment (Weisberg et al. 2009). The rapid uptake of teacher evaluation reforms came, in part, as a response to President Obama's RTTT program and its offer of large competitive grants to states that were struggling during the Great Recession (Bleiberg and Harbatkin 2020; Howell and Magazinnik 2017). In particular, the application rubric for RTTT rewarded states for using student outcomes to evaluate teachers and inform personnel decisions with evaluation ratings. Additionally, the Obama administration required states seeking a waiver from the No Child Left Behind

(NCLB) mandate to reach 100% proficiency by 2014 to commit to teacher evaluation reforms.

While states overwhelmingly adopted evaluation systems that broadly included the features rewarded by RTTT—weighting student outcomes, conducting annual evaluations, and using evaluation ratings for personnel decisions—there was substantial variation in the number and combination of adopted design features (Kraft and Gilmour 2017). The vast majority of states provided school districts with some degree of autonomy in designing and implementing teacher evaluation, either by allowing local discretion within a state-designed system or permitting districts to develop their systems given a set of guidelines (Steinberg and Donaldson 2016).

Mechanisms for Teacher Evaluation to Affect Student Outcomes

Theory predicts that performance evaluations are useful tools for improving worker output. Employers can use personnel evaluations to determine compensation and job responsibilities, as well as to provide feedback when objective measures are not available or cost-prohibitive (G. Baker 1992; Prendergast 1999). Data from performance management programs can provide helpful information to leaders of public sector organizations to improve outcomes (Heinrich 2002). In the K-12 education sector, there are two potential mechanisms through which teacher evaluation may impact student achievement and attainment. The first mechanism highlights the potential for evaluation reforms to change the composition of the teacher workforce by tying high-stakes personnel decisions such as dismissal and tenure decisions to performance ratings (Goldhaber and Hansen 2010; Gordon, Kane, and Staiger 2006; Staiger and Rockoff 2010; Liebowitz 2021; Sartain and Steinberg 2021). For example, several studies have found that new teacher evaluation systems increased voluntary turnover among lower-performing teachers (Cullen, Koedel, and Parsons 2021; Loeb, Miller, and Wyckoff 2015; Rodriguez, Swain, and Springer 2020; Steinberg

and Sartain 2015). Similarly, evidence from a national study of teacher evaluation reforms found that these reforms increased the number of new teaching candidates who had attended more competitive undergraduate institutions but also decreased the overall supply of teaching candidates (Kraft et al. 2020).

A second potential mechanism for improving student achievement and attainment via evaluation reforms is by improving current teachers' performance. Such improvements might reflect how the evaluation process promotes professional growth on the job and/or increased effort incentivized by dismissal threats or merit pay connected to evaluation scores (Donaldson and Papay 2015; Firestone 2014). The evaluation process itself may support ongoing improvements in teachers' practice if evaluators provide feedback and coaching, prompt teachers to reflect on their practices, or provide data that allow districts to match teachers with targeted professional development (Donaldson 2020; Donaldson and Firestone 2021; Galey-Horn and Woulfin 2021; Mintrop and Trujillo 2007; Springer 2010; Woulfin and Rigby 2017). Experimental studies of low-stakes observation and feedback by peers (Papay et al. 2020; Burgess, Rawal, and Taylor 2021) and administrators (Garet et al. 2017) have found some positive effects on achievement. However, field trials of training programs designed to improve evaluator feedback in high-stakes settings found no improvements on feedback quality or student achievement (Kraft and Christian in press; Mihaly et al. 2018), while a recent quasi-experimental study found no evidence that teachers alter their professional improvement activities in response to evaluation ratings (Koedel et al. 2019).

Evidence on the Effects of Teacher Evaluation Reforms on Student Achievement

Several quasi-experimental and experimental studies in large urban school districts point to the potential for evaluation systems to serve as engines for professional growth. Taylor and Tyler (2012) studied Cincinnati Public School's peer evaluation and feedback system. They found that being observed and

evaluated by experienced, expert teachers and school principals improved teachers' ability to raise student achievement in math but did not affect ELA achievement. A similar study of France's national teacher evaluation system found that high-stakes observation and feedback by certified pedagogical inspectors improved teachers' contributions to student achievement (Briole and Maurin 2020).

Research on the District of Columbia Public Schools' high-stakes teacher evaluation system, DC IMPACT, has found positive and sustained effects on student achievement (Dee, James, and Wyckoff 2021). The DC IMPACT system is unique in that it uses master educators and administrators as observers, places substantial weight on test-based measures of teacher performance, offers large financial incentives tied to performance ratings, and has resulted in the dismissal of a non-trivial number of teachers rated as low performing. Studies provide evidence that multiple mechanisms improved performance on the job for teachers (Dee and Wyckoff 2015; Phipps and Wiseman 2021) and teacher quality overall via selective retention and replacement (Adnot et al. 2016).

Evidence from studies examining teacher evaluation systems that are more representative of those adopted at scale nationally in the U.S. is decidedly mixed. In an experimental study of a pilot implementation of the new teacher evaluation system in Chicago Public Schools, Steinberg & Sartain (2015) found the pilot produced significant improvements in ELA achievement and positive but imprecisely estimated effects in math in the first year. However, the authors found no effect in either math or ELA among the cohort of schools that adopted the system in the second year, pointing to the challenges of sustaining effective evaluations at scale. An evaluation of the Gates Foundation's Intensive Partnerships for Effective Teaching, which provided $575 million to improve teacher evaluation across three large school districts and four charter management organizations, found that student achievement and graduation rates were largely

unchanged after five years (Stecher et al. 2018). Finally, a recent evaluation of a suite of teacher labor market reforms in Michigan, including teacher evaluation, reduced tenure protections, and reduced collective bargaining power, found largely null effects on student achievement (Anderson, Cowen, and Strunk 2019).

III. Data

Treatment

We draw on data from Kraft et al. (2020) to define the treatment timing of teacher evaluation reforms. We consider a state to be treated in the first year when districts were required to enact the new evaluation system statewide. Figure 1, Panel A, shows the 44 states that reformed teacher evaluation systems throughout the country. California, Iowa, Montana, Nebraska, Vermont, and Wyoming did not reform their teacher evaluation systems. Washington, D.C. was the first to reform its evaluation system, in 2009, while other states reformed teacher evaluation systems between 2012 to 2017 (See Appendix Figure A1).[3] The frequency of state reforms peaked in 2014 when 13 states reformed their teacher evaluation systems. The staggered timing in the rollout of reforms across states provides a unique opportunity to measure the effect of these evaluation systems on student outcomes.

<Insert Figure 1 Here>

We also collected data on 10 teacher evaluation policy components identified in the literature as key features of evaluation systems (Doherty and Jacobs 2015; Howell and Magazinnik 2017; NCTQ 2011; 2019). We then constructed an index equal to the number of teacher evaluation policy design

[3] Washington, D.C. does not contribute to the estimated effect of teacher evaluation on achievement because we do not observe pre-treatment math or ELA scores. We do observe pre-treatment attainment outcomes and leverage data from Washington, D.C. to identify those effects.

components that states required districts to put in place (See Appendix Table B1). As illustrated in Figure 1, Panel B, there was substantial variation in the design rigor of new evaluation systems across states (See Appendix Table B2 for state-specific data).

In addition to examining counts of policy components, we group the 10 design components into three categories based on their policy rationales (See Appendix Table B3). Sixteen states adopted a collection of reforms focused on enhancing the reliability of teacher evaluation measures, 19 adopted either incentives or accountability systems, and 29 used evaluations to provide feedback or inform professional development.

Outcomes

We use district-by-grade-level data from the Stanford Education Data Archive (SEDA), which includes a nearly complete census of school districts, to capture student achievement on high-stakes standardized state tests (Reardon et al. 2021).[4] The SEDA dataset links student performance across state-specific tests by norming scores relative to performance on the National Assessment of Education Progress (NAEP). SEDA includes test score estimates for third through eighth grade in math and ELA from 2009 to 2018.[5] Table 1 displays descriptive statistics for the full sample. We observe about 550,000 district-grade observations for both math and ELA.

<Insert Table 1 Here>

[4] In a few cases entire state-years are excluded from SEDA (Reardon et al. 2021). For example, if fewer than 95 percent of students took the state test or if multiple tests were administered for the same content area in the same year, then the entire state-year is excluded from SEDA.

[5] Test scores are aggregated in the SEDA up to the district-grade level and include all of the schools that fall within the borders of traditional public school districts.

To measure educational attainment, we construct state-by-year level estimates of high school graduation rates and college enrollment from the American Community Survey (ACS) Public Use Microdata Sample (PUMS). To measure high school graduation for each year and state, we calculate the proportion of 18-year-olds born in a state who earned a high school diploma or equivalent certificate relative to the total number of 18-year-olds born in a state, and apply appropriate PUMS person weights. To measure college enrollment for each state and year, we divide the number of 22-year-old students born in a state and enrolled in college in each year by the total number of 22-year-olds in a state and year, again using PUMS person weights from 2008 to 2019.[6] This procedure follows recent research on state education reforms that measures educational attainment based on the expected degree-earning age (Jackson, Wigger, and Xiong 2021; Rothstein and Schanzenbach 2021).[7]

Finally, we use the restricted NAEP student-level data on math and ELA achievement in fourth and eighth grades, available in odd-numbered years between 2003 to 2017, to replicate our core results. The NAEP assessment differs from the assessments used in SEDA in several relevant ways. First, the NAEP is not used for accountability purposes, removing any incentive for strategic behavior to increase scores. Second, the NAEP uses the same set of items for the entirety of the study period, improving the validity of comparisons across time, and measuring a broad range of competencies. Finally, the NAEP is limited in that it is administered only every other year and each assessment wave only includes a sample of approximately 4,000 schools (Sikali 2019).

Controls

[6] Appendix Table A1 describes the number of treated states and observations across relative time. The analytic sample is "trimmed" to mitigate weak panel balance.

[7] To avoid endogenous moves into states we use state of birth as a proxy for where a student attended school. Approximately 80 percent of students attend high school and college in their state of birth.

We supplement our main models with a parsimonious set of covariates. We add controls for the characteristics of schools and inputs to the educational production process related to student achievement or attainment. We measure all control variables prior to the first year of evaluation reforms and interact these baseline values with a time trend to control for potential differences in pre-treatment trends. This approach avoids including endogenous controls that may have been affected by the evaluation reforms themselves. In terms of school district characteristics, we include controls for district race and ethnicity (percent Black, percent Hispanic, percent Native American, and percent Asian), urbanicity, and total enrollment. Our education production process covariates include county level GDP, a poverty index, county unemployment rate, district-level student-teacher ratio, and district-level per-pupil expenditures.[8] We also add covariates for baseline outcomes to control for pre-treatment differences in student achievement and attainment. Data for the covariates from the achievement outcome models are from the SEDA 2.1 and 4.0 covariate files (Reardon et al. 2021). We obtain county-level GDP from the U.S. Bureau of Economic Analysis (2021) and district-level student/teacher ratios and per-pupil instructional expenditures from the Common Core of Data (U.S. Department of Education 2021). In the models with attainment outcomes, we use a parallel set of covariates measured at the state level from the NAEP, Common Core of Data, and Bureau of Economic Analysis.

IV. Method

We begin by fitting flexible event study models to test the parallel trends assumption and to explore the non-parametric evolution of any treatment effects:

[8] Poverty index is estimated using socioeconomic status proxies. For more details, see Reardon et al. (2021).

$$Y_{sdgt} = \sum_{k=-5}^{4} \tau_k 1(t = t_s^* + k) \times Tch\_Eval_s + \rho(\boldsymbol{X}'_{dt=2009} \times Year_t) + \alpha_d + \delta_g + \theta_t + \mu_{sdgt} \quad (1)$$

where Y_{sdgt} is a district-by-grade-by-year measure of mean achievement in grade g for district d in state s in year t (spring of school year). The term $1(t = t_s^* + k)$ represents a set of indicators for the years pre- and post-policy reform, with t_s^* denoting the year in which state s reformed its teacher evaluation system and $k \in$ [-5, 4]. Tch_Eval_{st} equals 1 for states that reformed teacher evaluation systems and zero otherwise. $\boldsymbol{X}$ is a vector of baseline covariates including the school district characteristics, education production process characteristics and baseline outcomes, discussed previously, all interacted with a linear time trend, $Year_t$. Each model also includes district fixed effects (α_d), grade fixed effects (δ_g), and year fixed effects (θ_t). The district fixed effects control for time-invariant district and state characteristics, including pre-treatment policies (e.g., standards-based reforms, teacher credentialing). The year and grade fixed effects control for year- and grade-specific shocks to achievement. μ is an idiosyncratic error term clustered at the state level.

The coefficients of primary interest in Equation 1 are the τ_k's, which represent the effect of teacher evaluation on our outcomes of interest k years before or after a reform. We measure these effects relative to the year just prior to the reform ($k = -1$) so that τ_{-3} and τ_1 represent the average effect of reforms on our outcomes of interest three years prior to and one year after reform, respectively.

To examine the non-parametric effect of teacher evaluation on educational attainment, we adapt Equation 1 to focus on our state-by-year measures. The state-level attainment models follow the same specification as the district-level

achievement models given by Equation 1, with a few differences. The baseline year in the attainment models is 2008 rather than 2009. The attainment models remove district and grade fixed effects, replacing them with state fixed effects. We also add baseline state-level controls (from 2008) for the percent of students eligible for free or reduced-price lunch (FRPL), percent Black, percent Hispanic, and average per-pupil expenditures, total student enrollment, NAEP scores, and the baseline outcome (either has a high school diploma or enrolled in college) all interacted with a linear time trend.

To improve precision, we complement our event studies with DiD specifications that take the following form:

$$Y_{sdgt} = \beta Tch\_Eval_{st} + \rho(\boldsymbol{X}'_{dt=2009} \times Year_t) + \alpha_d + \delta_g + \theta_t + \mu_{sdgt} \quad (2),$$

where Tch_Eval_{st} is an indicator that takes the value of unity if state *s* had enacted a teacher evaluation reform in year *t* and zero otherwise. All other variables are as defined in Equation 1. The coefficient of interest in Equation 2 is β, which is the DiD estimate of the effect of teacher evaluation averaged across the post-treatment years in our panel.

Our DiD framework relies on two key assumptions: 1) that comparison states provide a valid counterfactual for the trends in treated states in the absence of treatment; and 2) that there are no unobserved factors correlated with both our outcomes of interest and the timing of teacher evaluation reforms across states. We test the first assumption visually and empirically using the non-parametric event study. We also estimate a separate DiD model that includes state-specific linear time trends and examine the robustness of our results to the second assumption by fitting supplemental models that control for other education reforms that occurred within our panel window. The estimates from each approach are similar in sign and magnitude to those from our main DiD specification.

Several recent studies have shown that estimates from standard event studies and DiD specifications relying on the staggered timing of treatment for identification may be biased in the presence of heterogeneous treatment effects (Callaway and Sant'Anna 2020; Goodman-Bacon 2021; Sun and Abraham 2020). Consequently, we also report results from alternative TWFE estimators robust to issues related to heterogeneous treatment effects (A. Baker, Larcker, and Wang 2021; Cengiz et al. 2019; Sun and Abraham 2020). As we report below, our results are very consistent across these alternative estimation approaches.

V. Findings

Student Achievement

Event study estimates from models including baseline controls suggest that, on average, evaluation reforms did not affect students' performance in math or ELA. As shown in Figure 2, Panel A, in the first year of treatment (i.e., year 0), we can rule out positive effects as small as 0.003 SD for math and 0.005 SD for ELA.[9] Estimated effects in subsequent years are less precise, but even five years after treatment, we can rule out positive effects as small as 0.04 SD in both math and ELA. Our event study estimates also provide strong evidence that differential pre-trends do not drive our estimates: the pre-treatment estimates for all periods in math and ELA are individually and jointly indistinguishable from zero.

Our DiD estimates confirm these null effects and allow us to rule out small potential effects of teacher evaluation, averaged over all post-treatment years. Table 2, Panel A, includes the DiD estimates of the effect of teacher evaluation on student outcomes in math and ELA. The first column presents results without controls, and the second column includes baseline school, educational input, and achievement controls. After adding controls, we can rule out positive effects as small as 0.015 SD in math and about 0.009 SD in ELA.

[9] The event study estimates with and without controls are similar (see Appendix Table A2).

Educational Attainment

Similar to our achievement findings, event study and DiD estimates suggest that teacher evaluation had little effect on educational attainment. Figure 2, Panel B, provides the estimated effect of teacher evaluation on high school graduation and college enrollment. The effect of teacher evaluation on high school graduation and the percent enrolled in college are both small in magnitude and indistinguishable from zero. Importantly, we once again we find no evidence of differential pre-treatment trends. Estimates from event study models are precise enough to rule out a 2 percentage point increase in high school graduation and college enrollment across all observed years post-reform (See Appendix Table A3).

DiD results also show a null effect of teacher evaluation on education attainment. Table 2, Panel B, presents the DiD estimates for educational attainment, pooling over all post-treatment years. Our most precise estimates from models with covariates allow us to rule out a 1 percentage point increase in high school graduation and college enrollment.

<Insert Table 2 Here>

<Insert Figure 2 Here>

Heterogeneity by Evaluation System Design

Our average estimates may mask important treatment effect heterogeneity due to variation in system design. We test for potential heterogeneous effects across states based on our index of the number of design features a state required school districts to put in place. In Table 3, we present models that interact the

main treatment indicator with the continuous index of design rigor.[10] Overall, we find no evidence that high design rigor evaluation systems positively effected student achievement or attainment. The estimated coefficient on the interaction term between the treatment indicator and the design rigor index is statistically insignificant for three of our four outcomes. The one exception is ELA, where we find some evidence of negative differential effects. Specifically, states that required districts to put very few design components in place appear to have experienced small declines in student achievement post-reform. For example, our results in Table 3, Panel A, Column 4 suggest that in states which required districts to enact only two design features, the effect of teacher evaluation was -0.04 SD [95% CI: -0.07, -0.01].[11]

<Insert Table 3 Here>

The results in Table 3 suggest that, in general, the effect of teacher evaluation reform on student outcomes did not vary by the rigor of teacher evaluation system designs.[12] We provide further evidence of these null effects by plotting event studies for states with a high number of design components compared with states with a low number of design components separately. Figure 3 shows the event studies where the blue estimates are the effect of teacher evaluation for strong design states (i.e., systems with seven or more design features) relative to comparison states that did not adopt any reforms. The black estimates are the effect of teacher evaluation for states that adopted weaker

[10] In Table 3, the main effect of teacher evaluation is the effect of teacher evaluation for one state (i.e., Alabama) that implemented teacher evaluation, but did not choose a design that includes any of the components we observe in our index.
[11] The effect of enacting two design features is equal to the main effect of teacher evaluation plus the index multiplied by 2 (i.e., Evaluation+(2 X Index)).
[12] The results in Table 3 are similar when we use the first principal component from Principal Components Analysis.

designs (i.e., between one and six teacher evaluation design components) relative to comparison states. The effect of teacher evaluation is null for states with both stronger and weaker designs in math. Consistent with the differential effects by design rigor from Table 3, the event studies show small decreases in ELA scores one to two years after treatment for states with weak evaluation designs. As shown in Appendix Table A4, we find qualitatively similar results when estimating DiD models that pool across the post-treatment periods. The estimates with controls rule out positive effects as small as 0.039 SD in math, 0.040 SD in ELA, a 1 percentage point increase in high school students with diplomas, and a 2 percentage point increase in college enrollment for strong design states.[13]

<Insert Figure 3 Here>

Next, we use the 10 design components in our index to construct three non-mutually exclusive measures of specific policy rationales underlying teacher evaluation reforms: 1) reliable measurement; 2) incentives and accountability; and 3) professional development and feedback (see Appendix Table B2 for operationalizations of these dimensions and B3 for state counts). In Figure 4, Panel A, we plot event study estimates from states that adopted policy components to improve the reliability of teacher evaluation measures (e.g., use student test scores weighted at levels shown in research to yield reliable measures, at least two teaching observations, conduct student surveys). Figure 4, Panel B, plots estimates for states that tied incentives and accountability to teacher evaluation (e.g., bonuses, remove tenure). Figure 4, Panel C displays estimates for states that used teacher evaluation to inform professional development or provide

[13] Appendix Figure A2 mirrors Figure 2 except we change the definition of high quality to states that implemented eight or more teacher evaluation components. We find similarly precise null effects for states that implemented reforms with eight or more teacher evaluation components.

feedback to teachers. The blue line shows the effect for evaluation systems with a specific policy rationale, and the black line traces the effect of evaluation systems without the specified policy rationale. Overall, the event study estimates depicted in Figure 4 show little evidence that the effect of teacher evaluation reform varied with specific design components; the estimated coefficients are generally small in magnitude and statistically insignificant.

<Insert Figure 4 Here>

We conduct an additional heterogeneity analysis in attempt to better reconcile our consistent null results with prior research documenting the positive effects of evalution reforms in selected districts. In October of 2018, the National Council for Teacher Quality (NCTQ) released a report that profiled six district and state evaluation systems they judged to have designed and implemented exemplary evaluation systems that were producting results. These systems included Dallas Independent School District, Denver Public Schools, District of Columbia Public Schools, Newark Public Schools, Tennessee, and New Mexico (Putnam, Ross, and Walsh 2018). The report describes evidence that these exemplar systems successfully differentiated among teacher performance, retained higher-performing teachers and removed lower-performing teachers, and coincided with improvements in teacher evaluation ratings and student proficiency rates over time.

We test for differential effects among these exemplar systems by fitting models in which we disaggregate our indicator for treatment, Tch_Eval, into two mutually exclusive indicators identifying the implementation of new evaluation systems in 1) these exemplar districts and states and 2) all other states that adopted reforms (excluding the exemplary districts). Consistent with prior evidence, we find medium-sized positive effects of the implementation of these exemplar evaluation systems on math and ELA achievement. Figure 5 illustrates

both the null effects of evaluation among non-exemplary systems and the positive effects over time among exemplar systems rising to as high as 0.15 SD. In our pooled DiD model, we estimates a marginal significant positive effect of 0.09 SD in math and 0.07 SD effect in ELA (See Appendix Table A6).

We highlight two important caveats to these analyses. First, these exemplar districts were selected ex-post based on their outcomes by NCTQ making our findings somewhat unsurprising but not a forgone conclusion. We view this exemplar system heterogeneity analysis a helpful descriptive exercise that illustrates how precise null effects can mask positive effects among small subgroups. One might see this as encouraging confirmatory evidence that evaluation can work or discouraging given the difficulty of identify which systems will be successful ex-ante and why these systems appear to have moved the needle while the overwhelming majority did not. As NCTQ points out, "[these systems] are implementing many of the same components commonly found in many state districts systems" (Putnam, Ross, and Walsh 2018, 1). Second, these effects pool across two states and four districts but are largely driven by the state Tennessee of which represents 68% of the district-by-grade observations used to estimate the average effect of these exemplar systems.[14]

<Insert Figure 5 Here>

VI. Robustness Checks

Treatment Timing

We employ two alternative approaches to our standard event study models to test their robustness to potential heterogeneity across states and over time. We

[14] DCPS does not contribute to the estimated effects because no pre-treatment data is observed in SEDA for DCPS. We run a parallel set of models using NAEP that do include Washington, DC and find similar results.

first use a stacked DiD estimator (A. Baker, Larcker, and Wang 2021; Cengiz et al. 2019). To estimate the stacked event studies, we create six datasets, one for each cohort of states that reformed teacher evaluation systems in the same year (i.e., 2012, 2013, 2014, 2015, 2016, 2017), including the states in each cohort and the six states that never reformed their evaluation systems. We append the six datasets and supplement the models described in equations 1 and 2 by adding district-by-cohort and year-by-cohort fixed effects. Our second approach estimates cohort-specific average treatment effect on the treated (CATT) developed by Sun and Abraham (2020) . This approach is novel in that it calculates weights to estimate the CATT to correct the potential for negative weights in DiD event study models with staggered timing of adoption. Both approaches avoid identifying effects from comparing late to early reformers.

The null effects of teacher evaluation on achievement and attainment are robust to both estimation strategies that account for heterogeneous treatment effects across cohorts. Figure 6 includes event studies for each of the achievement and attainment outcomes from both alternative estimation approaches along with our main estimates. Across outcomes, the magnitude and sign of the estimates in each of the three models are quite similar. The effect of teacher evaluation across relative time remains insignificant. Together, these results suggest that our estimated null effects of teacher evaluation are not biased by treatment effect heterogeneity by adoption cohort.

<Insert Figure 6 Here>

Parallel Trends

The null effect of teacher evaluation is robust to the inclusion of state-specific linear trends, which provides additional evidence that the parallel trends assumption is met. Appendix Table A7 includes results for the achievement and attainment outcomes with and without covariates augmented with state-specific

linear trends.[15] The achievement results with state-specific linear trends are within 0.01 SD of the main results in Table 2. Similarly, the attainment results with state-specific linear trends differ by less than 1 percentage point from the main results. Overall, the effect of teacher evaluation remains insignificant after the inclusion of state-specific linear trends.

Contemporaneous Policies

Several other education policy reforms occurred contemporaneously during the period of adoption of teacher evaluation reforms. In particular, 17 states enacted reforms to teacher tenure between 2011 and 2014, with five eliminating tenure protections for new teachers and 12 increasing the number of probationary years for untenured teachers. Several states passed laws weakening collective bargaining for teachers between 2011 and 2016, with three restricting or eliminating mandatory collective bargaining and four eliminating mandatory union dues. Several states also enacted reforms to their school finance systems or adopted additional policies rewarded by RTTT (e.g., Common Core State Content Standards, school turnaround initiatives).[16]

Because these other reforms occurred in close temporal proximity to teacher evaluation reforms, they could bias our estimates of the impact of teacher evaluation reforms on student outcomes. To account for these potential confounding treatments, we specify models that add a vector of 19 time-varying education policies (Howell and Magazinnik 2017; Kraft et al. 2020). As shown in Appendix Table A8, we find similarly precise null effects for achievement and attainment outcomes after adding state policy controls. We can rule out positive effects as small as 0.01 SD for achievement outcomes and 1 percentage point for

[15] We present only effects without covariates for the attainment results because the state-level covariates interacted with the linear trends are collinear with the state-specific linear trends.
[16] See Kraft et al. (2020) for a complete listing of the education policy reforms that occurred contemporaneously during the sample timeframe.

attainment outcomes. The precisely estimated null effects suggest that unobserved education reforms do not bias the estimated effects of teacher evaluation.

Replicating Results in NAEP

We use the SEDA to measure student achievement in our preferred specification because it includes a near-census of school districts rather than a sampling of schools and is available every year rather than every other year. However, the state test scores used in the SEDA could reflect efforts to artificially raise scores due to the high-stakes attached to these tests (Ballou and Springer 2017; Booher-Jennings 2005; Neal and Schanzenbach 2010). To address this concern, we repeat our primary analyses using fourth- and eighth-grade math and ELA data from the low-stakes NAEP test. As shown in Appendix Table A9, consistent with our main results, we find null effects on achievement. We can rule out positive effects as small as 0.01 SD in math and 0.02 SD in ELA in models including controls.[17] These results add further support for our primary analyses using the SEDA.

VII. Extensions

Academically Vulnerable Groups

Advocates framed teacher evaluation reforms as essential to closing racial and socioeconomic achievement gaps (Weisberg et al. 2009). Consequently, in Appendix Table A10, we extend our primary analyses based on SEDA test scores to test for heterogeneity across sub-populations of students from different racial and socioeconomic backgrounds. Specifically, in our primary DiD specifications, we add interactions between the main effect of teacher evaluation and the percent of students in a district-grade-year eligible for FRPL, percent Black, and percent

[17] These models control for the same baseline district characteristics in Equation 1 and add student covariates, including sex, race/ethnicity, free or reduced lunch eligibility, limited English proficiency, has individualized education plan, and modal age for grade. We also add controls for state baseline math and ELA scores in 2003, and an indicator for whether a school made Adequate Yearly Progress in 2003 (Reback et al. 2013).

Hispanic measured at baseline. To improve the interpretability of estimates, we standardize each variable to have a mean of zero and a standard deviation of one. We find little evidence of heterogeneous effects. The estimated coefficient on the interaction between the treatment indicator and percent Hispanic for ELA and high school graduation is statistically significant and negative. This implies that, if anything, the reforms may have widened rather than closed achievement gaps between Hispanic and White students. However, the size of the effect is substantively small. The results in Appendix Table A10 suggest a 1 SD increase (20 percentage points) in the percent of Hispanic students leads to about a 0.03 SD decrease in ELA scores and a 0.3 percentage point decrease in high school graduation.

VIII. Conclusion

In this paper, we exploit the staggered timing of state teacher evaluation reforms to provide the first nationally representative evidence on how these reforms affected student achievement and educational attainment. We find that, on average, teacher evaluation reforms had no detectable effect on student achievement or attainment. We also find little evidence that the effect of teacher evaluation reforms varied depending on design rigor of the new evaluation systems states implemented or that teacher evaluation improved outcomes for the academically vulnerable groups it was intended to benefit. These null effects are robust to a wide range of specification checks, including alternative TWFE estimators, the inclusion of state-specific linear trends, and controlling for other contemporaneous education reforms.

As noted previously, several school districts have demonstrated success with teacher evaluation reforms, including Cincinnati (Taylor and Tyler 2012), Chicago (Steinberg and Sartain 2015; Sartain and Steinberg 2021), and Washington, DC (Adnot et al. 2016; Dee and Wyckoff 2015; Dotter, Chaplin, and Bartlett 2021; James and Wyckoff 2020). We affirm the precense of exemplary

systems contribute to our national analyses by documenting the positive effects of a small set of systems identified as exemplary ex-post. This leads naturally to the question of why, at the national level, teacher evaluation reforms appear to have had little impact on student outcomes.

While we cannot provide a definitive answer to that question, we believe part of the answer is tied to the disconnect between the best practices for performance management systems and the actual design and implementation of new state systems. Despite the widespread adoption of teacher evaluation reforms, many states designed evaluation systems that only vaguely resembled the systems most reformers envisioned. The federal government used RTTT and NCLB waivers to influence the design of new teacher evaluation systems (Howell and Magazinnik 2017), but this influence had its limits. For example, only 19 states adopted design features intended to link high-stakes accountability and incentives to performance ratings and only 16 established rigorous multi-measure evaluation systems.

Even when states adopted more rigorous design features, these features were rarely sustained over time or implemented in ways that resembled the high-stakes systems shown to have positive effects in prior research (NCTQ 2019). Such systems appear to have been organizationally, economically, and politically challenging to scale across a diverse and decentralized U.S. public education system. For example, nationally, less than one percent of teachers were rated as unsatisfactory under the new evaluation systems, with performance-based dismissals being exceedingly rare (Kraft and Gilmour 2017). Similarly, states that did link evaluation to compensation often offered small bonuses of only a few hundred to a thousand dollars and set the bar so low that most teachers qualified for the bonuses (NCTQ 2019). As a result, the accountability components of teacher evaluation systems were often designed and implemented in ways that rendered them low-stakes (Aldeman and Chuong 2014).

Evidence also suggests that evaluation reforms were sometimes implemented in ways that resulted in unintended consequences, a further possible explanation for our null results and the small negative effects we find in some contexts. Prior research documents that teacher evaluation reforms decreased job satisfaction and perceived autonomy among new teachers (Kraft et al. 2020). New evaluation systems created large demands on administrators' time to conduct frequent observations and complete considerable paperwork, displacing other more potentially productive activities (Neumerski et al. 2018). Many districts also placed unrealistic expectations on administrators to provide critical feedback to teachers, narrowing the scope, depth, and quality of feedback teachers received (Kraft and Christian in press; Hunter and Springer in press).

Firms in the private sector often fail to implement best management practices and performance evaluation systems because of imperfectly competitive markets and the costs of implementing such policies and practices (Bloom and Van Reenen 2007). These same factors are likely to have influenced the design and implementation of teacher evaluation reforms. Unlike firms in a perfectly competitive market with incentives to implement management and evaluation systems that increase productivity, school districts and states face less competitive pressure to innovate. Similarly, adopting evaluation systems like the one implemented in Washington D.C. requires a significant investment of time, money, and political capital. Many states may have believed that the costs of these investments outweighed the benefits. Consequently, the evaluation systems adopted by many states were not meaningfully different from the status quo and subsequently failed to improve student outcomes.

Reference List

Adnot, Melinda, Thomas Dee, Veronica Katz, and James Wyckoff. 2016. "Teacher Turnover, Teacher Quality, and Student Achievement in Dcps." *Educational Evaluation and Policy Analysis*, 0162373716663646.

Aldeman, Chad, and Carolyn Chuong. 2014. "Teacher Evaluations in an Era of Rapid Change: From" Unsatisfactory" to" Needs Improvement"." *Bellwether Education Partners*.

Anderson, Kaitlin P., Joshua M. Cowen, and Katharine O. Strunk. 2019. "The Impact of Teacher Labor Market Reforms on Student Achievement: Evidence from Michigan." *Education Finance and Policy*, 1–43.

Baker, Andrew, David F. Larcker, and Charles CY Wang. 2021. "How Much Should We Trust Staggered Difference-In-Differences Estimates?" *Available at SSRN 3794018*.

Baker, George. 1992. "Incentive Contracts and Performance Measurement." *Journal of Political Economy* 100 (3): 598–614.

Ballou, Dale, and Matthew G. Springer. 2017. "Has NCLB Encouraged Educational Triage? Accountability and the Distribution of Achievement Gains." *Education Finance and Policy* 12 (1): 77–106.

Bleiberg, Joshua, and Erica Harbatkin. 2020. "Teacher Evaluation Reform: A Convergence of Federal and Local Forces." *Educational Policy* 34 (6): 918–52.

Bloom, Nicholas, and John Van Reenen. 2007. "Measuring and Explaining Management Practices across Firms and Countries." *The Quarterly Journal of Economics* 122 (4): 1351–1408.

———. 2011. "Human Resource Management and Productivity." In *Handbook of Labor Economics*, 4:1697–1767. Elsevier.

Booher-Jennings, Jennifer. 2005. "Below the Bubble:'Educational Triage' and the Texas Accountability System." *American Educational Research Journal* 42 (2): 231–68.

Briole, Simon, and Éric Maurin. 2020. "There's Always Room for Improvement: The Persistent Benefits of Repeated Teacher Evaluations." In .

Burgess, Simon, Shenila Rawal, and Eric S. Taylor. 2021. "Teacher Peer Observation and Student Test Scores: Evidence from a Field Experiment in English Secondary Schools." *Journal of Labor Economics* 39 (4): 1155–86. https://doi.org/10.1086/712997.

Callaway, Brantly, and Pedro HC Sant'Anna. 2020. "Difference-in-Differences with Multiple Time Periods." *Journal of Econometrics*.

Cappelli, Peter, and Martin J. Conyon. 2018. "What Do Performance Appraisals Do?" *ILR Review* 71 (1): 88–116.

Cengiz, Doruk, Arindrajit Dube, Attila Lindner, and Ben Zipperer. 2019. "The Effect of Minimum Wages on Low-Wage Jobs." *The Quarterly Journal of Economics* 134 (3): 1405–54.

Chambers, J., I. Brodziak de los Reyes, and C. O'Neil. 2013. "How Much Are Districts Spending to Implement Teacher Evaluation Systems." RAND.

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

Coburn, Cynthia E. 2003. "Rethinking Scale: Moving beyond Numbers to Deep and Lasting Change." *Educational Researcher* 32 (6): 3–12.

Cullen, Julie Berry, Cory Koedel, and Eric Parsons. 2021. "The Compositional Effect of Rigorous Teacher Evaluation on Workforce Quality." *Education Finance and Policy* 16 (1): 7–41.

Dee, Thomas S., Jessalynn James, and Jim Wyckoff. 2021. "Is Effective Teacher Evaluation Sustainable? Evidence from District of Columbia Public Schools." *Education Finance and Policy* 16 (2): 313–46.

Dee, Thomas S., and James Wyckoff. 2015. "Incentives, Selection, and Teacher Performance: Evidence from IMPACT." *Journal of Policy Analysis and Management* 34 (2): 267–97.

Doherty, Kathryn, and Sandi Jacobs. 2015. "State of the States 2015: Evaluating Teaching, Leading and Learning." *National Council on Teacher Quality*. National Council on Teacher Quality.

Donaldson, Morgaen. 2020. *Multidisciplinary Perspectives on Teacher Evaluation: Understanding the Research and Theory*. Routledge.

Donaldson, Morgaen, and William Firestone. 2021. "Rethinking Teacher Evaluation Using Human, Social, and Material Capital." *Journal of Educational Change*, 1–34.

Donaldson, Morgaen, and John Papay. 2015. "Teacher Evaluation for Accountability and Development." In *Handbook of Research in Education Finance and Policy*, 2nd ed. Routledge.

Dotter, Dallas, Duncan Chaplin, and Maria Bartlett. 2021. "Impacts of School Reforms in Washington, DC on Student Achievement." Mathematica Policy Research.

Firestone, William A. 2014. "Teacher Evaluation Policy and Conflicting Theories of Motivation." *Educational Researcher* 43 (2): 100–107.

Galey-Horn, Sarah, and Sarah l Woulfin. 2021. "Muddy Waters: The Micropolitics of Instructional Coaches' Work in Evaluation." *American Journal of Education* 127 (3): 000–000.

Garet, Michael S., Andrew J. Wayne, Seth Brown, Jordan Rickles, Mengli Song, and David Manzeske. 2017. "The Impact of Providing Performance

Feedback to Teachers and Principals. NCEE 2018-4001." *National Center for Education Evaluation and Regional Assistance*.
Gibbons, Robert. 1998. "Incentives in Organizations." *Journal of Economic Perspectives* 12 (4): 115–32.
Goldhaber, Dan, and Michael Hansen. 2010. "Using Performance on the Job to Inform Teacher Tenure Decisions." *American Economic Review* 100 (2): 250–55.
Goodman-Bacon, Andrew. 2021. "Difference-in-Differences with Variation in Treatment Timing." *Journal of Econometrics*.
Gordon, Robert James, Thomas J. Kane, and Douglas Staiger. 2006. In *Identifying Effective Teachers Using Performance on the Job*. Brookings Institution Washington, DC.
Government Accountability Office. 2015. "Race to the Top: Survey of State Education Agencies' Capacity to Implement Reform (GAO-15-316SP, April 2015), an E-Supplement to GAO-15-295." Washington, D.C. http://www.gao.gov/products/gao-15-316sp.
Gupta, Snigdha, Lauren H. Supplee, Dana Suskind, and John A. List. 2021. "Failed to Scale: Embracing the Challenge of Scaling in Early Childhood." In *The Scale-Up Effect in Early Childhood and Public Policy*, 1–21. Routledge.
Hanushek, Eric A. 2009. "Teacher Deselection." In *Creating a New Teaching Profession*, edited by Dan Goldhaber and Jane Hanaway, 165–80. Urban Institute.
Heinrich, Carolyn J. 2002. "Outcomes–Based Performance Management in the Public Sector: Implications for Government Accountability and Effectiveness." *Public Administration Review* 62 (6): 712–25.
Heinrich, Carolyn J., and Gerald Marschke. 2010. "Incentives and Their Dynamics in Public Sector Performance Management Systems." *Journal of Policy Analysis and Management* 29 (1): 183–208.
Heinrich, Carolyn J., Robert H. Meyer, and Greg Whitten. 2010. "Supplemental Education Services under No Child Left Behind: Who Signs up, and What Do They Gain?" *Educational Evaluation and Policy Analysis* 32 (2): 273–98.
Honig, Meredith I. 2006. *New Directions in Education Policy Implementation: Confronting Complexity*. Suny Press.
Howell, William G., and Asya Magazinnik. 2017. "Presidential Prescriptions for State Policy: Obama's Race to the Top Initiative." *Journal of Policy Analysis and Management* 36 (3): 502–31.
Hunter, Seth B, and Matthew G. Springer. in press. "Critical Feedback Characteristics, Teacher Human Capital, and Early-Career Teacher Performance: A Mixed-Methods Analysis Using Written Feedback from

Formal Evaluation Conferences." *Educational Evaluation and Policy Analysis*.

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

Jackson, C. Kirabo, Cora Wigger, and Heyu Xiong. 2021. "Do School Spending Cuts Matter? Evidence from the Great Recession." *American Economic Journal: Economic Policy* 13 (2): 304–35.

James, Jessalynn, and James H. Wyckoff. 2020. "Teacher Evaluation and Teacher Turnover in Equilibrium: Evidence from DC Public Schools." *AERA Open* 6 (2): 2332858420932235.

Koedel, Cory, Jiaxi Li, Matthew G. Springer, and Li Tan. 2019. "Teacher Performance Ratings and Professional Improvement." *Journal of Research on Educational Effectiveness* 12 (1): 90–115.

Kraft, Matthew. 2019. "Teacher Effects on Complex Cognitive Skills and Social-Emotional Competencies." *Journal of Human Resources* 54 (1): 1–36.

Kraft, Matthew, Eric J. Brunner, Shaun M. Dougherty, and David J. Schwegman. 2020. "Teacher Accountability Reforms and the Supply and Quality of New Teachers." *Journal of Public Economics* 188: 104212.

Kraft, Matthew, and Alvin Christian. in press. "Can Teacher Evaluation Systems Produce High-Quality Feedback? An Administrator Training Field Experiment." *American Educational Research Journal*.

Kraft, Matthew, and Allison Gilmour. 2017. "Revisiting the Widget Effect: Teacher Evaluation Reforms and the Distribution of Teacher Effectiveness." *Educational Researcher* 46 (5): 234–49.

Liebowitz, David D. 2021. "Teacher Evaluation for Accountability and Growth: Should Policy Treat Them as Complements or Substitutes?" *Labour Economics*, 102024.

Loeb, Susanna, Luke C. Miller, and James Wyckoff. 2015. "Performance Screens for School Improvement: The Case of Teacher Tenure Reform in New York City." *Educational Researcher* 44 (4): 199–212.

Manna, Paul. 2010. *Collision Course: Federal Education Policy Meets State and Local Realities*. Washington, D.C.: CQ Press.

McGuinn, Patrick. 2012. "Stimulating Reform: Race to the Top, Competitive Grants and the Obama Education Agenda." *Educational Policy* 26 (1).

Mihaly, Kata, Heather L. Schwartz, Isaac M. Opper, Geoffrey Grimm, Luis Rodriguez, and Louis T. Mariano. 2018. "Impact of a Checklist on Principal-Teacher Feedback Conferences Following Classroom Observations. REL 2018-285." *Regional Educational Laboratory Southwest*.

Mintrop, Heinrich, and Tina Trujillo. 2007. "The Practical Relevance of Accountability Systems for School Improvement: A Descriptive Analysis of California Schools." *Educational Evaluation and Policy Analysis* 29 (4): 319–52.

Murphy, Joseph, Philip Hallinger, and Ronald H. Heck. 2013. "Leading via Teacher Evaluation: The Case of the Missing Clothes?" *Educational Researcher* 42 (6): 349–54.

NCTQ. 2011. "State of the States 2011: Trends and Early Lessons on Teacher Evaluation and Effectiveness Policies." *National Council on Teacher Quality*.

———. 2019. "Teacher & Principal Evaluation Policy. State of the States 2019." *National Council on Teacher Quality*.

Neal, Derek, and Diane Whitmore Schanzenbach. 2010. "Left behind by Design: Proficiency Counts and Test-Based Accountability." *The Review of Economics and Statistics* 92 (2): 263–83.

Neumerski, Christine M., Jason A. Grissom, Ellen Goldring, Mollie Rubin, Marisa Cannata, Patrick Schuermann, and Timothy A. Drake. 2018. "Restructuring Instructional Leadership: How Multiple-Measure Teacher Evaluation Systems Are Redefining the Role of the School Principal." *The Elementary School Journal* 119 (2): 270–97.

Oyer, P., and S. Schaefer. 2011. "Personnel Economics: Hiring and Incentives." In *Handbook of Labor Economics*. Vol. 4. Elsevier.

Papay, John P., Eric S. Taylor, John H. Tyler, and Mary E. Laski. 2020. "Learning Job Skills from Colleagues at Work: Evidence from a Field Experiment Using Teacher Performance Data." *American Economic Journal: Economic Policy* 12 (1): 359–88.

Petek, Nathan, and Nolan Pope. 2016. "The Multidimensional Impact of Teachers on Students." University of Chicago Working Paper.

Phipps, Aaron R., and Emily A. Wiseman. 2021. "Enacting the Rubric: Teacher Improvements in Windows of High-Stakes Observation." *Education Finance and Policy* 16 (2): 283–312.

Prendergast, Canice. 1999. "The Provision of Incentives in Firms." *Journal of Economic Literature* 37 (1): 7–63.

Pressman, Jeffrey L., and Aaron Wildavsky. 1984. *Implementation: How Great Expectations in Washington Are Dashed in Oakland; Or, Why It's Amazing That Federal Programs Work at All, This Being a Saga of the Economic Development Administration as Told by Two Sympathetic Observers Who Seek to Build Morals on a Foundation*. Vol. 708. Univ of California Press.

Putnam, Hannah, Elizabeth Ross, and Kate Walsh. 2018. “Making a Difference: Six Places Where Teacher Evaluation Systems Are Getting Results.” National Council on Teacher Quality.

Reardon, Sean, Andrew Ho, Benjamin Shear, Erin Fahle, Demetra Kalogrides, and Belen Chavez. 2021. “Stanford Education Data Archive (Version 4.0).” http://purl.stanford.edu/db586ns4974.

Rodriguez, Luis A., Walker A. Swain, and Matthew G. Springer. 2020. “Sorting through Performance Evaluations: The Influence of Performance Evaluation Reform on Teacher Attrition and Mobility.” *American Educational Research Journal* 57 (6): 2339–77.

Rothstein, Jesse, and Diane Whitmore Schanzenbach. 2021. “Does Money Still Matter? Attainment and Earnings Effects of Post-1990 School Finance Reforms.” National Bureau of Economic Research.

Sartain, Lauren, and Matthew P. Steinberg. 2016. “Teachers’ Labor Market Responses to Performance Evaluation Reform: Experimental Evidence from Chicago Public Schools.” *Journal of Human Resources* 51 (3): 615–55.

———. 2021. “Can Personnel Policy Improve Teacher Quality? The Role of Evaluation and the Impact of Exiting Low-Performing Teachers.” *EdWorkingPapers.Com*. Annenberg Institute at Brown University.

Sawchuk, Stephen. 2015. “Teacher Evaluation Heads to the Courts.” Education Week. October 7, 2015. https://www.edweek.org/policy-politics/teacher-evaluation-heads-to-the-courts.

Sikali, Emmanuel. 2019. “NAEP 2017 National and State Mathematics and Reading, and Puerto Rico Mathematics (Grades 4 & 8) Restricted-Use Data Files.” NCES.

Springer, Matthew G. 2010. *Performance Incentives: Their Growing Impact on American K-12 Education*. Brookings Institution Press.

Staiger, Douglas O., and Jonah E. Rockoff. 2010. “Searching for Effective Teachers with Imperfect Information.” *Journal of Economic Perspectives* 24 (3): 97–118.

Stecher, Brian M., Deborah J. Holtzman, Michael S. Garet, Laura S. Hamilton, John Engberg, Elizabeth D. Steiner, Abby Robyn, et al. 2018. “Improving Teaching Effectiveness: Final Report: The Intensive Partnerships for Effective Teaching Through 2015–2016,” June.

Steinberg, Matthew P., and Morgaen Donaldson. 2016. “The New Educational Accountability: Understanding the Landscape of Teacher Evaluation in the Post-NCLB Era.” *Education Finance and Policy*.

Steinberg, Matthew P., and Lauren Sartain. 2015. “Does Teacher Evaluation Improve School Performance? Experimental Evidence from Chicago’s Excellence in Teaching Project.” *Education Finance and Policy*.

Sun, Liyang, and Sarah Abraham. 2020. "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects." *Journal of Econometrics*.

Taylor, Eric S., and John H. Tyler. 2012. "The Effect of Evaluation on Teacher Performance." *The American Economic Review* 102 (7): 3628–51.

U.S. Bureau of Economic Analysis. 2021. "CAGDP2 Gross Domestic Product (GDP) by County and Metropolitan Area." bea.gov.

U.S. Department of Education. 2021. "Common Core of Data." https://nces.ed.gov/ccd/ccddata.asp.

Weisberg, Daniel, Susan Sexton, Jennifer Mulhern, and David Keeling. 2009. "The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness. Executive Summary." *New Teacher Project*.

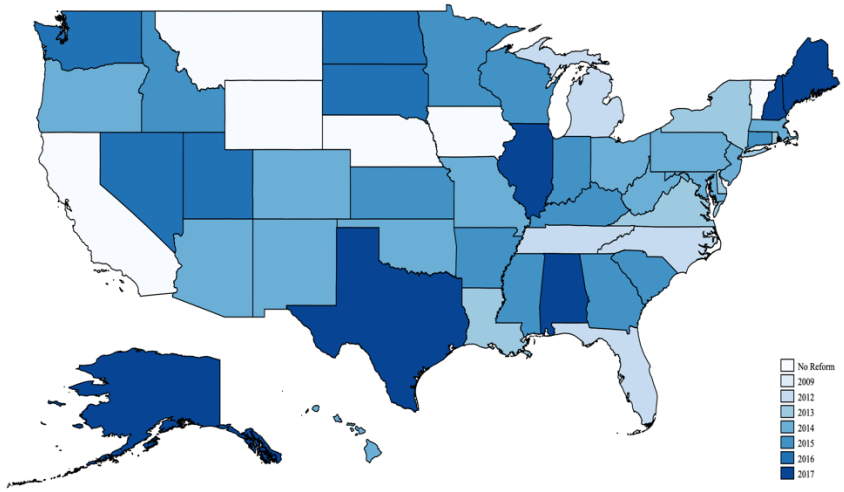
Woulfin, Sarah L., and Jessica G. Rigby. 2017. "Coaching for Coherence: How Instructional Coaches Lead Change in the Evaluation Era." *Educational Researcher* 46 (6): 323–28.

Zhou, Jin, Alison Baulos, James J. Heckman, and Bei Liu. 2021. "The Economics of Investing in Early Childhood." *The Scale-Up Effect in Early Childhood and Public Policy: Why Interventions Lose Impact at Scale and What We Can Do About It*.

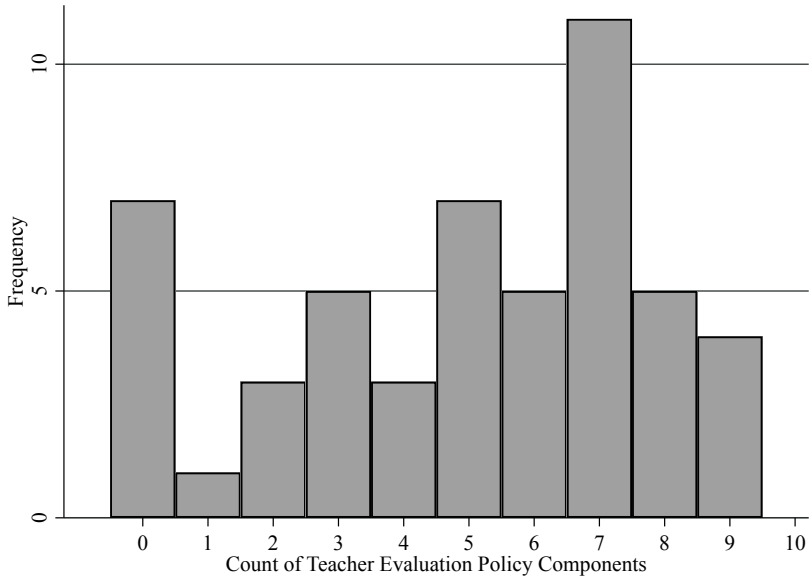
Zhou, Jin, Alison Baulos, James J. Heckman, and Bei Liu. 2021. "The Economics of Investing in Early Childhood: Importance of Understanding the Science of Scaling." In The Scale-Up Effect in Early Childhood and Public Policy (pp. 76-97). Routledge.

Figure 1. Teacher Evaluation Implementation

Panel A. State Implementation Map



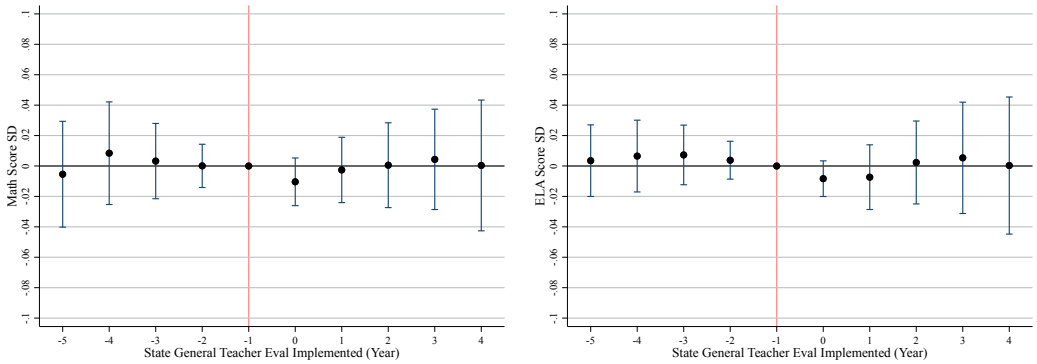
Panel B. Histogram of Teacher Evaluation Reform Quality Index



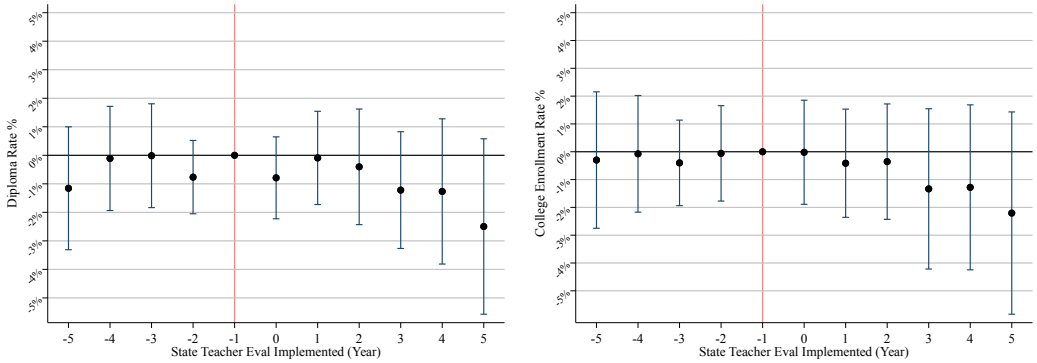
Note: The index for comparison states is zero even if they implemented a component of teacher evaluation reform. See Appendix B for details on the components of the index. All years are the spring of the school year.

Figure 2. Event Study: Effects on Achievement and Attainment

Panel A. Achievement



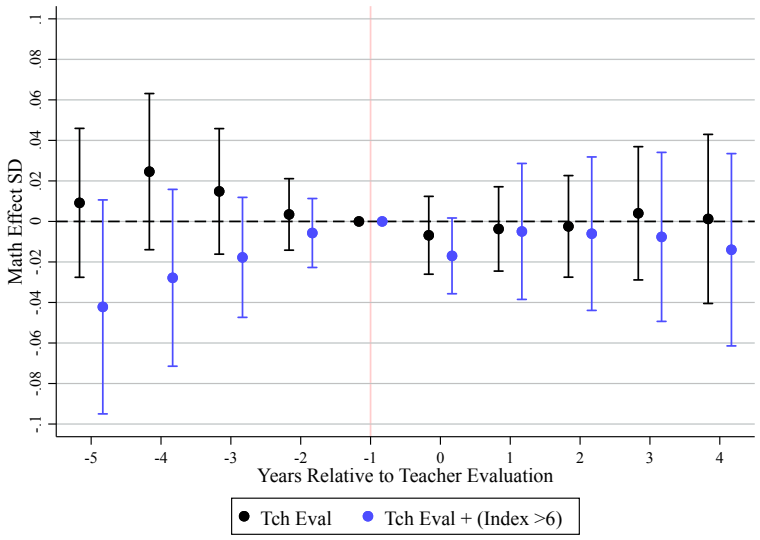
Panel B. Attainment



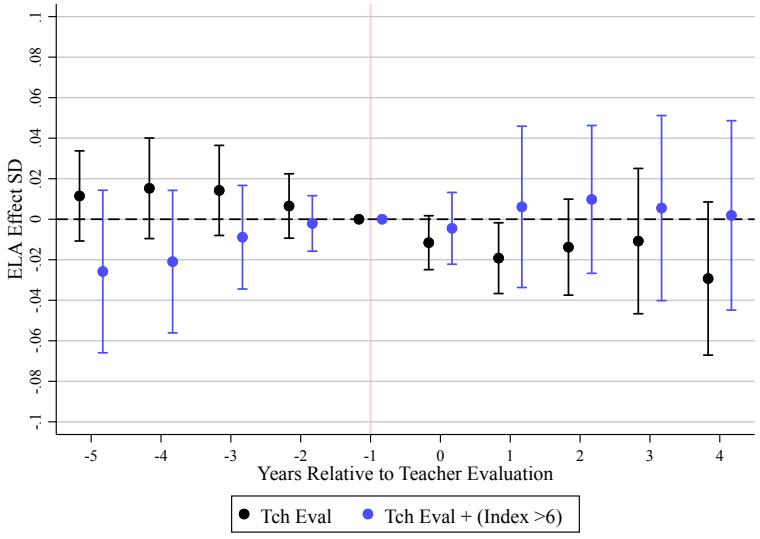
Note: Models with achievement outcomes include district fixed effects, grade fixed effects, year fixed effects, and baseline covariates measured in 2009 interacted with a linear year trend: percent Black, percent Hispanic, percent Native American, percent Asian, total enrollment, urban/city, GDP, poverty index, unemployment rate, student teacher ratio, per-pupil expenditures, ELA score, and math score. Models with attainment outcomes include state fixed effects, year fixed effects, and baseline covariates measured in 2009 interacted with a linear year trend: percent Black, percent Hispanic, percent Native American, percent Asian, total enrollment, urban/city, GDP, percent FRPL, unemployment rate, student teacher ratio, per-pupil expenditures, and either baseline high school graduation or baseline college enrollment. Standard errors are clustered by state.

Figure 3. Event Study: Heterogeneity by Index

Panel A. Math



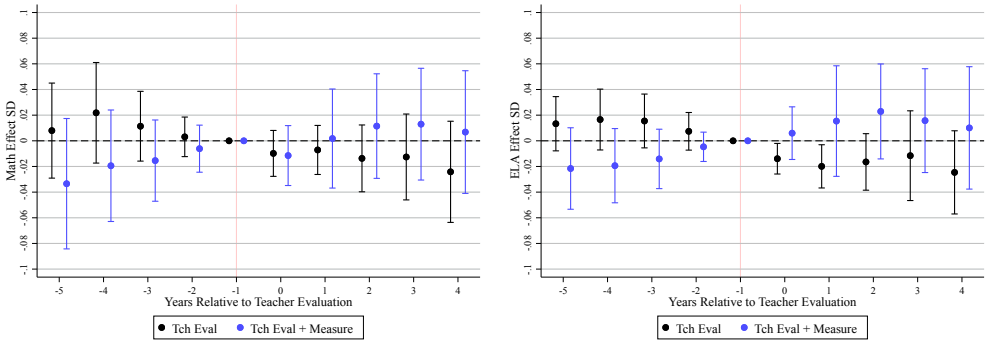
Panel B. ELA



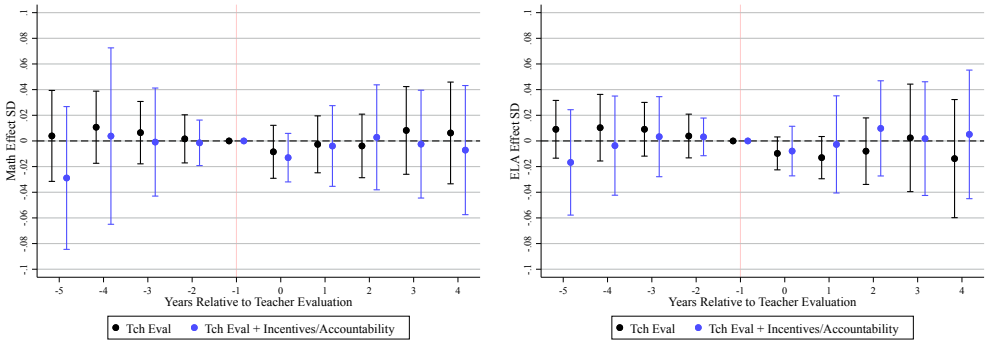
Note: Models include two estimates for each relative time period: the main event study dummies and a set of event study dummies interacted with a time-invariant indicator equal to one for states that had an index from 7 to 10. The black estimates "Tch Eval" are the main event study dummies. The blue estimates "Tch Eval + (Index >6)" are the linear combination of the estimates for the "high group" estimates and main event study estimate from the same relative time period. 20 states have an index from 7 to 10. Model specification found in notes for Figure 2. Standard errors are clustered by state.

Figure 4. Event Study: Heterogeneity by System Design

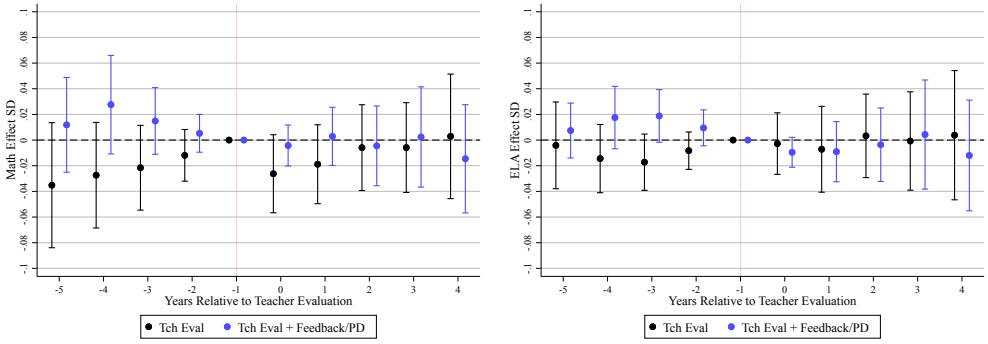
Panel A. Measurement



Panel B. Accountability and Incentives



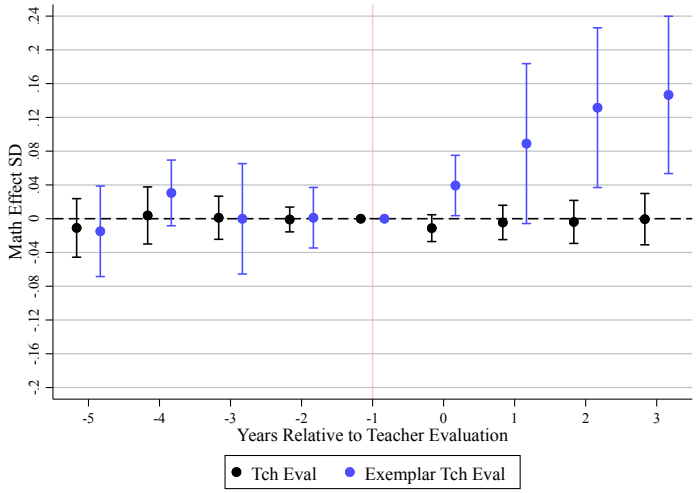
Panel B. Feedback and Professional Development



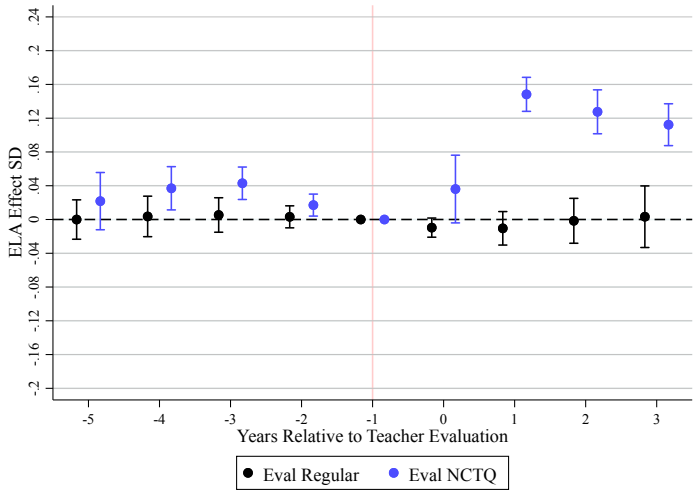
Note: Models include the main event study dummies and a set of event study dummies interacted with a time-invariant indicator equal to one for a specified system design. The black estimates are the main event study dummies. The blue estimates are the linear combination of the estimates for the "high group" estimates and main event study estimate from the same relative time period. See Table B3 for state system design details. Model specification found in notes for Figure 2. Standard errors are clustered by state.

Figure 5. Event Study: Heterogeneity by Exemplar Evaluation Systems

Panel A. Math

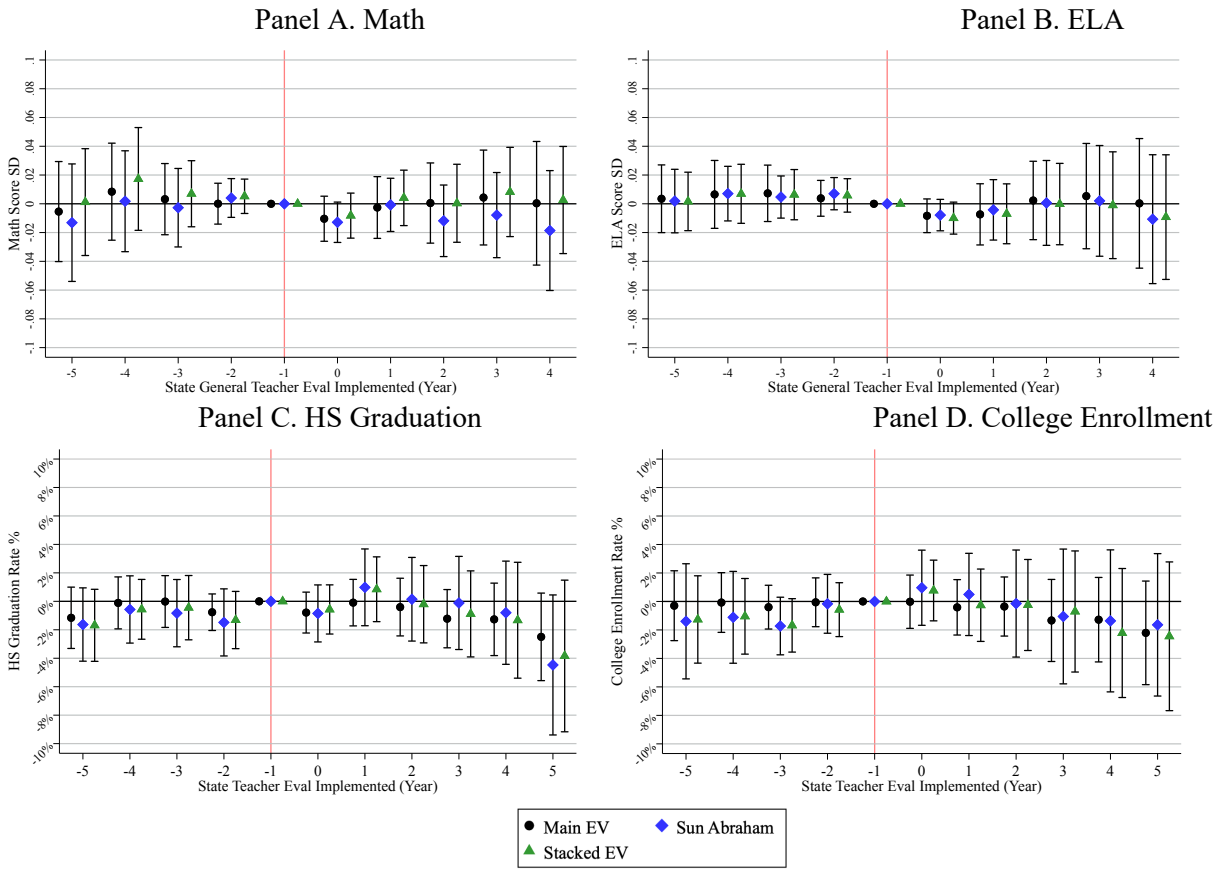


Panel B. ELA



Note: Exemplar districts/states are Dallas Independent School District, Denver Public Schools, Newark Public Schools, Tennessee, and New Mexico. Models include two estimates for each relative time period: the main event study dummies and a set of event study dummies interacted with a time-invariant indicator equal to one for states that were exemplar districts. We present effects up to 4 years after adoption of evaluation systems because Tennessee is the only exemplar system we observe outcomes for 5 years after treatment. The black estimates "Tch Eval" are the main event study dummies. The blue estimates "Exemplar Tch Eval" are the effect of teacher evaluation for the exemplar districts and states. Model specification found in notes for Figure 2. Standard errors are clustered by state.

Figure 6. Event Study and Estimaes Robust to Heterogenous Effects Across Cohorts



Note: Model specification found in notes for Figure 2. Main event study duplicates the results from Figures 2. Diamonds indicate CATT estimates and triangle are stacked event study estimates. Each model includes six stacks with a cohort of treated states and six never treated states. Standard errors are clustered by state by stack.

Table 1. Analytic Sample Descriptive Characteristics

Characteristic	Mean	SD	N	Source
ELA Score	0.041	0.38	491,944	SEDA
Math Score	0.041	0.41	460,401	SEDA
High School Graduation	61.3	6.48	520	ACS
College Enrollment	62.6	6.07	520	ACS
Percent White	0.74	0.27	460,287	SEDA
Percent Black	0.08	0.17	460,287	SEDA
Percent Hispanic/Latinx	0.13	0.20	460,287	SEDA
Percent Native American	0.03	0.10	460,287	SEDA
Percent Asian	0.02	0.05	460,287	SEDA
Total Enrollment (Ks)	327.17	979.73	460,287	SEDA
Urban/City	0.07	0.26	460,287	SEDA
GDP Chained $s (100Ks)	23.55	68.24	460,401	BEA
Poverty Index	0.13	0.07	460,287	SEDA
Unemployment Rate	0.07	0.03	460,287	SEDA
Student Teacher Ratio	15.12	4.16	447,509	CCD
Per-Pupil Expenditures in Ks	6.768	3.60	459,906	CCD

Note: SEDA=Stanford Education Data Archive; ACS=American Community Survey; BEA=Bureaus of Economic Analysis; CCD=Common Core of Data. Table 1 includes descriptive statistics for units included in the analytic sample from the regressions for each outcome. Covariate descriptive restricted are estimated using the district data from the SEDA math sample.

Table 2. Effect of Teacher Evaluation: Difference-in-Differences Models

	(1)	(2)	(3)	(4)
	Panel A. Achievement			
Outcome	Math	Math	ELA	ELA
Teacher Evaluation	-0.0184	-0.0080	-0.0220	-0.0098
	(0.0126)	(0.0115)	(0.0120)	(0.0096)
District FE	X	X	X	X
Grade FE	X	X	X	X
Year FE	X	X	X	X
District Ed Controls		X		X
Local SES Controls		X		X
Achievement Controls		X		X
n	460,401	460,401	491,944	491,944
	Panel B. Attainment			
Outcome	HS Grad	HS Grad	College Enroll	College Enroll
Teacher Evaluation	0.3996	-0.0718	0.0885	0.0860
	(0.6677)	(0.6363)	(0.6785)	(0.7002)
State FE	X	X	X	X
Year FE	X	X	X	X
State Ed Controls		X		X
State SES Controls		X		X
Attainment Controls		X		X
n	520	520	520	520

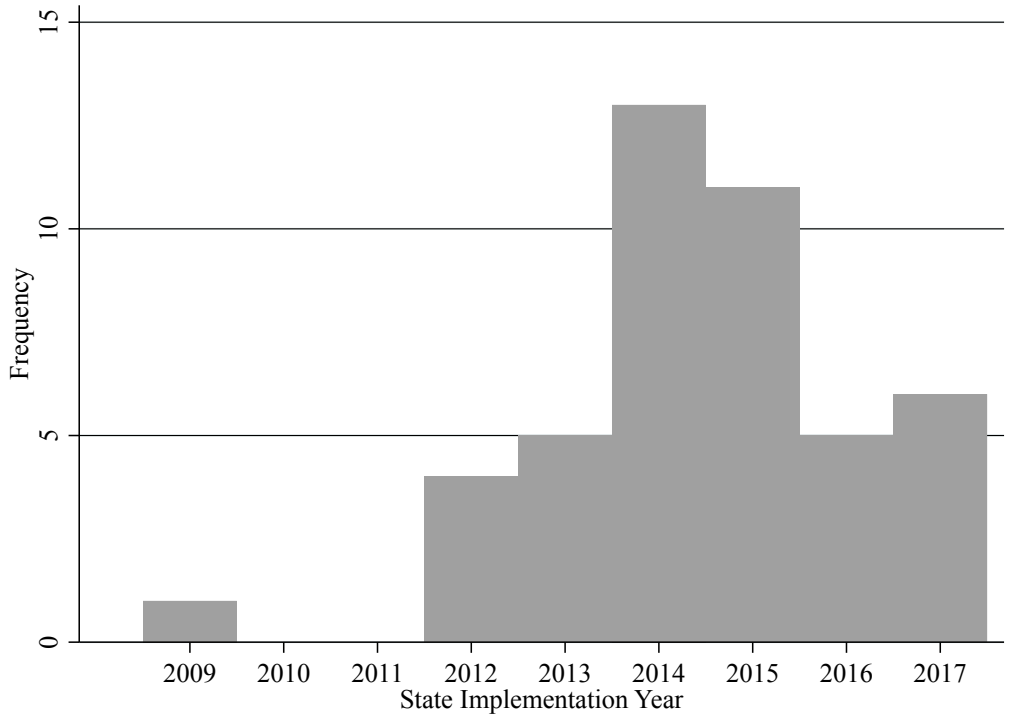
Note: Models with achievement outcomes include district fixed effects, grade fixed effects, year fixed effects, and baseline covariates measured in 2009 interacted with a linear year trend: Percent Black, Percent Hispanic, Percent Native American, Percent Asian, Total Enrollment, Urban/City, GDP, Poverty Index, Unemployment Rate, Student Teacher Ratio, Per-Pupil Expenditures, ELA Score, and Math Score. Models with attainment outcomes include state fixed effects, year fixed effects, and baseline covariates measured in 2009 interacted with a linear year trend: Percent Black, Percent Hispanic, Percent Native American, Percent Asian, Total Enrollment, Urban/City, GDP, Percent FRPL, Unemployment Rate, Student Teacher Ratio, Per-Pupil Expenditures, either baseline High School Graduation or Baseline College Enrollment. Standard errors are clustered by state.

Table 3. Regressing Continuous Teacher Quality Index on Outcomes

	(1)	(2)	(3)	(4)
	Panel A. Achievement			
Outcome	Math	Math	ELA	ELA
Teacher Evaluation	-0.0474	-0.0318	-0.0686	-0.0590
	(0.0275)	(0.0248)	(0.0214)	(0.0217)
Teacher Evaluation X Index	0.0052	0.0043	0.0085	0.0090
	(0.0048)	(0.0044)	(0.0037)	(0.0039)
District FE	X	X	X	X
Year FE	X	X	X	X
Grade FE	X	X	X	X
District Ed Controls		X		X
Local SES Controls		X		X
Achievement Controls		X		X
n	460,401	460,401	491,944	491,944
	Panel B. Attainment			
Outcome	HS Grad	HS Grad	College Enroll	College Enroll
Teacher Evaluation	1.5430	1.3745	-0.3799	-0.8878
	(0.9753)	(0.8877)	(1.0460)	(0.9822)
Teacher Evaluation X Index	-0.2035	-0.2520	0.0834	0.1689
	(0.1508)	(0.1406)	(0.1721)	(0.1677)
State FE	X	X	X	X
Year FE	X	X	X	X
State Ed Controls		X		X
State SES Controls		X		X
Attainment Controls		X		X
n	520	520	520	520

Note: See notes in Table 2 for a full list of covariates. Standard errors are clustered by state. The main effect of teacher evaluation is the effect of teacher evaluation for one state (i.e., Alabama) that implemented teacher evaluation, but did not choose a design that includes any of the components we observe in our index.

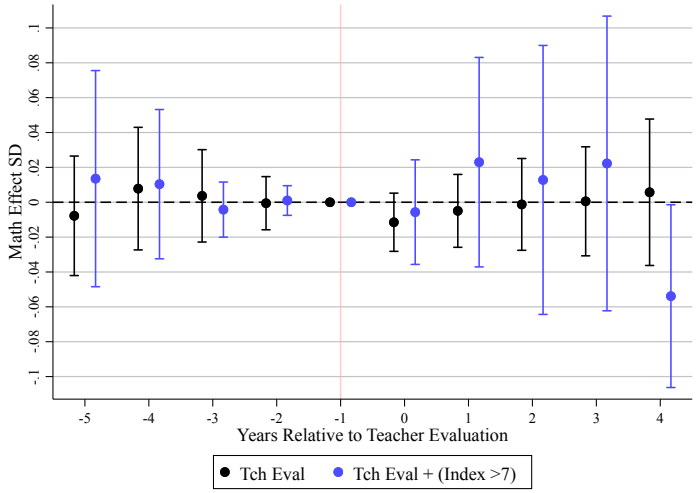
Appendix Figure A1. State Implementation of Teacher Evaluation Reforms by Year



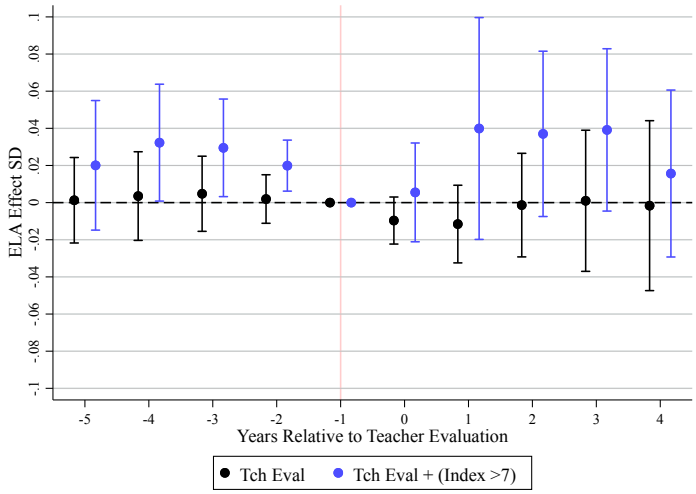
Note: All years are the spring of the school year.

Appendix Figure A2. Event Study Rigorous Design (Index 8 to 10)

Panel A. Math



Panel B. ELA



Note: Models include two estimates for each relative time period: the main event study dummies and a set of event study dummies interacted with a time-invariant indicator equal to one for states that had an index from 8 to 10. The black estimates "Tch Eval" are the main event study dummies. The blue estimates "Tch Eval + (Index > 7)" are the linear combination of the estimates for the "high group" estimates and main event study estimate from the same relative time period. 9 states have an index from 8 to 11: CT, DC, DE, GA, LA, NJ, RI, TN, and UT. Model specification found in notes for Figure 2. Standard errors are clustered by state.

Appendix Table A1. Observations and States Across Relative Time

Relative Time	Treated States	N	Trimmed
	Panel A. Achievement		
Pre -8	6	22,896	X
Pre -7	11	28,133	X
Pre -6	22	54,788	X
Pre -5	34	88,631	
Pre -4	39	99,255	
Pre -3	43	108,143	
Pre -2	41	102,721	
Pre -1	38	97,970	
Post 0	41	94,711	
Post 1	41	88,263	
Post 2	38	72,867	
Post 3	33	70,386	
Post 4	19	41,662	
Post 5	9	11,308	X
Post 6	5	9,361	X
	Panel B. Attainment		
Pre -8	6	6	X
Pre -7	11	11	X
Pre -6	22	22	X
Pre -5	35	35	
Pre -4	40	40	
Pre -3	44	44	
Pre -2	44	44	
Pre -1	44	44	
Post 0	45	45	
Post 1	44	44	
Post 2	44	44	
Post 3	44	44	
Post 4	38	38	
Post 5	33	33	
Post 6	22	22	X
Post 7	9	9	X
Post 8	4	4	X

Note: Treated states indicates the number of treated states observable for a specified relative time period. For achievement outcomes, N indicates the number of district-grade observations pooled across subject for a specified relative time period. For attainment outcomes the unit of analysis is state so the unique number of states and number of observations is identical.

Appendix Table A2. Event Study Achievement Effects

	(1)	(2)
	Panel A. Math	
-5 Pre	0.0215	-0.0054
	(0.0165)	(0.0173)
-4 Pre	0.0256	0.0084
	(0.0162)	(0.0168)
-3 Pre	0.0150	0.0032
	(0.0123)	(0.0123)
-2 Pre	0.0054	0.0001
	(0.0073)	(0.0071)
0 Post	-0.0157	-0.0104
	(0.0081)	(0.0078)
1 Post	-0.0137	-0.0026
	(0.0111)	(0.0107)
2 Post	-0.0187	0.0005
	(0.0145)	(0.0139)
3 Post	-0.0205	0.0044
	(0.0181)	(0.0164)
4 Post	-0.0248	0.0004
	(0.0233)	(0.0214)
District FE	X	X
Grade FE	X	X
Year	X	X
District Ed Controls		X
Local SES Controls		X
Achievement Controls		X
n	460,401	460,401
	Panel B. ELA	
-5 Pre	0.0235	0.0035
	(0.0146)	(0.0117)
-4 Pre	0.0222	0.0065
	(0.0136)	(0.0118)
-3 Pre	0.0183	0.0073
	(0.0107)	(0.0098)
-2 Pre	0.0093	0.0038
	(0.0067)	(0.0062)
0 Post	-0.0144	-0.0083
	(0.0062)	(0.0058)
1 Post	-0.0193	-0.0073
	(0.0113)	(0.0106)
2 Post	-0.0161	0.0023
	(0.0152)	(0.0136)
3 Post	-0.0203	0.0054
	(0.0209)	(0.0182)
4 Post	-0.0315	0.0003
	(0.0261)	(0.0224)
n	491,944	491,944

Note: See notes in Table 2 for a full list of covariates. Model 1 includes state and year fixed effects. Model 2 adds district education, SES, and achievement controls. Standard errors are clustered by state.

Appendix Table A3. Event Study Attainment Effects

	(1)	(2)
	Panel A. HS Graduation	
-5 Pre	-1.9309	-1.1564
	(1.1552)	(1.0726)
-4 Pre	-0.7242	-0.1103
	(1.0252)	(0.9090)
-3 Pre	-0.4345	-0.0142
	(0.9235)	(0.9061)
-2 Pre	-0.9788	-0.7651
	(0.6393)	(0.6396)
0 Post	-0.5336	-0.7897
	(0.7072)	(0.7150)
1 Post	0.4394	-0.0916
	(0.8247)	(0.8138)
2 Post	0.4157	-0.4042
	(0.9060)	(1.0090)
3 Post	-0.0831	-1.2201
	(0.9239)	(1.0187)
4 Post	0.2264	-1.2667
	(1.0375)	(1.2675)
5 Post	-0.6339	-2.4965
	(1.0944)	(1.5312)
n	520	520
	Panel B. College Enrollment	
-5 Pre	-0.5615	-0.2984
	(1.0750)	(1.2211)
-4 Pre	-0.3048	-0.0749
	(0.9649)	(1.0438)
-3 Pre	-0.5730	-0.4004
	(0.6965)	(0.7653)
-2 Pre	-0.1636	-0.0582
	(0.8210)	(0.8535)
0 Post	0.0430	-0.0185
	(0.9159)	(0.9316)
1 Post	-0.2863	-0.4124
	(0.9525)	(0.9685)
2 Post	-0.1475	-0.3558
	(0.9568)	(1.0330)
3 Post	-1.0473	-1.3349
	(1.3214)	(1.4345)
4 Post	-0.8688	-1.2801
	(1.3485)	(1.4766)
5 Post	-1.6587	-2.2053
	(1.5795)	(1.8105)
n	520	520

Note: See notes in Table 2 for a full list of covariates. Model 1 includes state and year fixed effects. Model 2 adds state covariates and attainment controls. Standard errors are clustered by state.

Appendix Table A4. Effect of Rigorously Designed Teacher Evaluation

	(1)	(2)	(3)	(4)
	Panel A. Achievement			
Outcome	Math	Math	ELA	ELA
Teacher Evaluation	-0.0273	-0.0161	-0.0361	-0.0248
	(0.0148)	(0.0135)	(0.0133)	(0.0109)
Eval X High Quality	0.0221	0.0203	0.0357	0.0385
	(0.0208)	(0.0201)	(0.0177)	(0.0169)
District FE	X	X	X	X
Year FE	X	X	X	X
Grade FE	X	X	X	X
District Ed Controls		X		X
Local SES Controls		X		X
Achievement Controls		X		X
n	460,401	460,401	491,944	491,944
	Panel B. Attainment			
Outcome	HS Grad	HS Grad	College Enroll	College Enroll
Teacher Evaluation	0.4409	0.0914	0.1002	0.0601
	(0.7144)	(0.6527)	(0.6657)	(0.6558)
Eval X High Quality	-0.0924	-0.3787	-0.0261	0.0587
	(0.7450)	(0.7039)	(0.7286)	(0.7211)
State FE	X	X	X	X
Year FE	X	X	X	X
State Ed Controls		X		X
State SES Controls		X		X
Attainment Controls		X		X
n	520	520	520	520

Note: See notes in Table 2 for a full list of covariates. High quality indicates an index value of 7, 8, or 9. For full list of covariates see Table 1. Standard errors are clustered by state.

Appendix Table A5. Moderation Analysis with Theoretical Constructs

	(1)	(2)	(3)	(4)	(5)	(6)
			Panel A. Achievement			
Outcome	Math	Math	Math	ELA	ELA	ELA
Eval	-0.0206	-0.0109	0.0069	-0.0261	-0.0177	0.0073
	(0.0127)	(0.0137)	(0.0187)	(0.0099)	(0.0101)	(0.0165)
Eval X Measurement	0.0396			0.0533		
	(0.0184)			(0.0170)		
Eval X Incent/Account		0.0064			0.0180	
		(0.0206)			(0.0181)	
Eval X Feedback/PD			-0.0219			-0.0254
			(0.0198)			(0.0188)
District FE	X	X	X	X	X	X
Year FE	X	X	X	X	X	X
Grade FE	X	X	X	X	X	X
District Ed Controls	X	X	X	X	X	X
Local SES Controls	X	X	X	X	X	X
Achievement Controls	X	X	X	X	X	X
n	460,401	460,401	460,401	491,944	491,944	491,944
			Panel B. Attainment			
Outcome	HS Grad	HS Grad	HS Grad	College Enroll	College Enroll	College Enroll
Teacher Evaluation	-0.0141	0.0380	0.0881	0.2515	0.2240	0.7133
	(0.2112)	(0.2189)	(0.3036)	(0.5912)	(0.7161)	(0.7188)
Eval X Measurement	0.0443			-0.8687		
	(0.2273)			(0.8514)		
Eval X Incent/Account		-0.0828			-0.7212	
		(0.2195)			(0.6925)	
Eval X Feedback/PD			-0.1234			-1.1436
			(0.2594)			(0.7160)
State FE	X	X	X	X	X	X
Year FE	X	X	X	X	X	X
State Ed Controls	X	X	X	X	X	X
State SES Controls	X	X	X	X	X	X
Attainment Controls	X	X	X	X	X	X
n	520	520	520	520	520	520

Note: See notes in Table 2 for a full list of covariates. Each model includes all fixed effects and controls. See Appendix Table B3 for a full list of states that belong to each construct.

Appendix Table A6. Effects of Exemplar Evaluation Systems

	(1)	(2)	(3)	(4)
Outcome	Math	Math	ELA	ELA
Teacher Evaluation	-0.0220	-0.0105	-0.0120	-0.0120
	(0.0125)	(0.0114)	(0.0094)	(0.0094)
Exemplar	0.0855	0.0925	0.0702	0.0702
	(0.0549)	(0.0527)	(0.0296)	(0.0296)
District FE	X	X	X	X
Year FE	X	X	X	X
Grade FE	X	X	X	X
District Ed Controls		X		X
Local SES Controls		X		X
Achievement Controls		X		X
n	440,565	440,565	471,797	471,797

Note: Exemplar districts/states are Dallas Independent School District, Denver Public Schools, Newark Public Schools, Tennessee, and New Mexico. Models include two estimates for each relative time period: the main event study dummies and a set of event study dummies interacted with a time-invariant indicator equal to one for states that were exemplar districts. We present effects up to 4 years after adoption of evaluation systems because Tennessee is the only exemplar system we observe outcomes for 5 years after treatment. The black estimates "Tch Eval" are the main event study dummies. The blue estimates "Exemplar Tch Eval" are the effect of teacher evaluation for the exemplar districts and states. Model specification found in notes for Figure 2. Standard errors are clustered by state.

Appendix Table A7. Controlling for State-Specific Linear Trends

	(1)	(2)
	Panel A. Achievement	
Outcome	Math	Math
Teacher Evaluation	-0.0044	-0.0044
	(0.0125)	(0.0125)
District FE	X	X
Grade FE	X	X
Year FE	X	X
District Ed Controls		X
Local SES Controls		X
Achievement Controls		X
State-Specific Trends	X	X
n	460,401	460,401
Outcome	ELA	ELA
Teacher Evaluation	-0.0044	-0.0043
	(0.0080)	(0.0080)
District FE	X	X
Grade FE	X	X
Year FE	X	X
District Ed Controls		X
Local SES Controls		X
Achievement Controls		X
State-Specific Trends	X	X
n	491,944	491,944
	Panel B. Attainment	
Outcome	HS Grad	College Enroll
Teacher Evaluation	-0.1086	0.2126
	(0.8974)	(1.0088)
State FE	X	X
Year FE	X	X
State-Specific Trends	X	X
n	520	520

Note: See notes in Table 2 for a full list of covariates. Standard errors are clustered by state. Covariates in the models with attainment outcomes are interacted with linear time trends and are perfectly collinear with the state-specific trends.

Appendix Table A8. Controlling for Time Varying State Policies

	(1)	(2)	(3)	(4)
	Panel A. Achievement			
Outcome	Math	Math	ELA	ELA
Teacher Evaluation	-0.0123	-0.0033	-0.0206	-0.0077
	(0.0136)	(0.0106)	(0.0138)	(0.0099)
District FE	X	X	X	X
Grade FE	X	X	X	X
Year FE	X	X	X	X
District Ed Controls		X		X
Local SES Controls		X		X
Achievement Controls		X		X
State Policies	X	X	X	X
n	455,388	455,388	486,663	486,663
	Panel B. Attainment			
Outcome	HS Grad	HS Grad	College Enroll	College Enroll
Teacher Evaluation	0.5227	0.1217	-0.5513	-0.5141
	(0.6655)	(0.6573)	(0.6321)	(0.6298)
State FE	X	X	X	X
Year FE	X	X	X	X
State Ed Controls		X		X
State SES Controls		X		X
Attainment Controls		X		X
State Policies	X	X	X	X
n	513	513	513	513

Note: See notes in Table 2 for a full list of covariates. Standard errors are clustered by state. Policy covariates from Kraft et al (2020) and Howell & Magazinnik (2017) include eliminate tenure, increase probationary period, weaken collective bargaining, eliminate mandatory union dues, won Race to the Top, implement Common Core, basic skills licensure tests, content area licensure tests, pedagogical knowledge licensure tests, Common Core assessment, charter authorizer, charter building funds, charter cap, school turnaround, alternative teacher certification, vouchers, high school exit exams, summative testing, and school finance reform interacted with state quartiles of median household income (2000).

Appendix Table A9. Replicating results using the Low-Stakes NAEP Assessment

	(1)	(2)
	Panel A. Math	
Teacher Evaluation	-0.063	-0.024
	(0.026)	(0.013)
District FE	X	X
Grade FE	X	X
Year FE	X	X
Student Controls		X
District Ed Controls		X
Achievement Controls		X
n	1,480,590	1,480,590
	Panel B. ELA	
Teacher Evaluation	-0.058	0.002
	(0.044)	(0.011)
District FE	X	X
Grade FE	X	X
Year FE	X	X
Student Controls		X
District Ed Controls		X
Achievement Controls		X
n	1,397,020	1,397,020

Note: Student covariates include sex, race/ethnicity, Free and Reduced Price Lunch Eligibility, Limited English Proficiency, has Individualized Education Plan, and modal age for grade. District covariates includes all the district characteristics included in Table 1. NAEP samples sizes rounded in accordance with NCES restricted use rules. Achievement characteristics include state baseline math and ELA scores in 2003 and a school level indicator of whether a school made Adequate Yearly Progress. NAEP results use student-level inverse probability weights. Standard errors are clustered by state.

Appendix Table A10. Differential Effects for Sub-Groups

	(1)	(2)	(3)
Panel A. Math			
Teacher Evaluation	-0.0125	-0.0103	-0.0140
	(0.0120)	(0.0117)	(0.0123)
Teacher Evaluation X Percent FRPL	-0.0138		
	(0.0092)		
Teacher Evaluation X Percent Black		-0.0007	
		(0.0052)	
Teacher Evaluation X Percent Hispanic			-0.0108
			(0.0056)
n	450,163	450,163	450,163
Panel B. ELA			
Teacher Eval	-0.0210	-0.0201	-0.0152
	(0.0129)	(0.0127)	(0.0090)
Teacher Evaluation X Percent FRPL	-0.0069		
	(0.0068)		
Teacher Evaluation X Percent Black		-0.0019	
		(0.0052)	
Teacher Evaluation X Percent Hispanic			-0.0179
			(0.0051)
n	480,801	480,801	480,801
Panel C. High School Graduation			
Teacher Eval	-0.1558	-0.1179	-0.0908
	(0.6235)	(0.6235)	(0.6279)
Teacher Evaluation X Percent FRPL	-1.0503		
	(0.6475)		
Teacher Evaluation X Percent Black		-0.4382	
		(0.7577)	
Teacher Evaluation X Percent Hispanic			-0.7385
			(0.5881)
n	520	520	520
Panel D. College Enrollment			
Teacher Eval	0.1392	0.1937	0.0897
	(0.7012)	(0.7116)	(0.6985)
Teacher Evaluation X Percent FRPL	0.7934		
	(0.8557)		
Teacher Evaluation X Percent Black		1.0920	
		(0.7692)	
Teacher Evaluation X Percent Hispanic			0.5846
			(0.6704)
n	520	520	520

Note: Models with achievement outcomes includes district, year, and grade fixed effects, district education controls, local SES controls, and achievement controls. Models with attainment outcomes include state, year fixed effects, state education, state SES controls, and attainment controls. See notes in Table 2 for a full list of covariates. Standard errors are clustered by state. Poverty rate, percent Black, and percent Hispanic are all measured at baseline (2009) and standardized.

Appendix Table B1. Teacher Evaluation Reform Components

Category	Variable	Descriptions	Source	State #
Accountability/ Incentive	Fire Teachers	Tenured and untenured teachers rated "ineffective" may be removed from their position.	Howell & Magazinnik (2017)	28
Accountability/ Incentive	Grant Tenure	Teacher evaluation ratings used to grant tenure and/or full certification.	Howell & Magazinnik (2017)	29
Accountability/ Incentive	Bonus	Providing additional compensation to teachers rated "highly effective".	Howell & Magazinnik (2017)	20
Accountability/ Incentive	Career Ladder	Providing additional responsibilities to teachers rated "highly effective".	Howell & Magazinnik (2017)	11
Measurement	Multiple Categories	Evaluations have three or more rating categories.	Howell & Magazinnik (2017)	38
Measurement	Observations Required	Observations are a required component of teacher evaluations.	Doherty & Jacobs (2015)	27
Measurement	Student Survey	Student surveys are a required component of teacher evaluations.	Doherty & Jacobs (2015)	7
Measurement	Student data	Student test scores (e.g., growth scores, value-added) with a weight of 20-50 percent are a required component of teacher evaluations.	Bleiberg & Harbatkin (2020); Doherty & Jacobs (2015)	21
Feedback/PD	Feedback Required	Teachers receive feedback based on their evaluations.	Doherty & Jacobs (2015)	35
Feedback/PD	Inform PD	Teacher evaluations inform coaching, induction support, and/or professional development.	Howell & Magazinnik (2017)	36

Note: Howell & Magazinnik (2017) do not include data for DC. Design features for DC were determined using the NCTQ State of the State reports from three years were used were used (Doherty and Jacobs 2015; NCTQ 2011; 2019).

Appendix Table B2. Teacher Evaluation Categorical Constructs and Quality Measures

Category	Descriptions	State #
Measurement	Teacher evaluation systems include at least three of the following components: (1) Student test scores weighted 20 to 50 percent; (2) observations [at least two explicitly required]; (3) student surveys; (4) Evaluations have three or more rating categories.	16
Accountability/ Incentive	Teacher evaluation systems include at least three of the following components: (1) Evaluation used to either grant tenure or (2) remove teachers from their position and evaluations used for either (3) promotions or (4) bonuses.	19
Feedback/PD	Teachers must receive feedback based on their evaluation; have their evaluation inform coaching, induction support and/or professional development.	29
Low Quality	State index value is 0 to 3.	10
Medium Quality	State index value is 4 to 6.	15
High Quality	State index value is 7 to 9.	20

Appendix Table B3. State Teacher Evaluation Component Measures by State

State	Ever Adopted	Measurement	Accountability/ Incentive	Feedback/PD	Index
AK	1	0	0	0	4
AL	1	0	0	0	0
AR	1	0	1	1	7
AZ	1	0	0	1	5
CA	0	0	0	0	0
CO	1	0	1	1	7
CT	1	1	1	1	9
DC	1	1	1	0	8
DE	1	0	1	1	7
FL	1	0	1	1	7
GA	1	1	1	1	9
HI	1	1	0	1	8
IA	0	0	0	0	0
ID	1	0	0	0	3
IL	1	0	0	1	6
IN	1	1	1	0	7
KS	1	0	0	0	4
KY	1	1	0	1	6
LA	1	1	1	1	8
MA	1	0	1	1	8
MD	1	0	0	0	3
ME	1	1	0	1	7
MI	1	0	1	1	7
MN	1	0	0	0	3
MO	1	0	0	1	3
MS	1	0	0	0	1
MT	0	0	0	0	0
NC	1	0	0	0	5
ND	1	0	0	0	2
NE	0	0	0	0	0
NH	1	0	1	1	6
NJ	1	1	0	1	7
NM	1	1	0	1	5
NV	1	0	1	1	7
NY	1	0	1	1	6
OH	1	1	1	0	7
OK	1	1	1	0	7
OR	1	0	0	0	3
PA	1	1	0	0	5
RI	1	1	1	1	9
SC	1	0	0	0	2
SD	1	0	0	1	5
TN	1	1	1	1	8
TX	1	0	0	1	2
UT	1	1	1	1	9
VA	1	0	0	0	4
VT	0	0	0	0	0
WA	1	0	0	1	5
WI	1	0	0	1	6
WV	1	0	0	1	5
WY	0	0	0	0	0