NBER WORKING PAPER SERIES

TAKING TEACHER EVALUATION TO SCALE: THE EFFECT OF STATE REFORMS ON ACHIEVEMENT AND ATTAINMENT

Joshua Bleiberg
Eric Brunner
Erica Harbatkin
Matthew A. Kraft
Matthew G. Springer

Working Paper 30995
http://www.nber.org/papers/w30995

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2023

Corresponding author, Joshua Bleiberg, can be reached at 230 South Bouquet Street Pittsburgh, PA 15260 or jbleiber@pitt.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 NBER Economics of Education Program Meetings, the Brown Half-Baked Research Series, the Annual Northeast Economics of Education Workshop, and the Society for Research on Educational effectiveness annual conference. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Taking Teacher Evaluation to Scale: The Effect of State Reforms on Achievement and Attainment
Joshua Bleiberg, Eric Brunner, Erica Harbatkin, Matthew A. Kraft, and Matthew G. Springer
NBER Working Paper No. 30995
March 2023
JEL No. I20,I21,I28

ABSTRACT

Federal incentives and requirements under the Obama administration spurred states to adopt major reforms to their 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 0.015 standard deviation for achievement and 1 percentage point for high school graduation and college enrollment. We also find little evidence that the effect of teacher evaluation reforms varied by system design rigor, specific design features or student and district characteristics. We highlight five factors that may have undercut the efficacy of teacher evaluation reforms at scale: political opposition, the decentralized structure of U.S. public education, capacity constraints, limited generalizability, and the lack of increased teacher compensation to offset the non-pecuniary costs of lower job satisfaction and security.

Joshua Bleiberg
University of Pittsburgh
jbleiber@pitt.edu

Eric Brunner
Department of Public Policy
University of Connecticut 10
Prospect Street, 4th Floor
Hartford, CT 06103
eric.brunner1@gmail.com

Erica Harbatkin
Florida State University
Tallahassee, FL 32306
eharbatkin@fsu.edu

Matthew A. Kraft
Brown University
PO Box 1938
Providence, RI 02478
and NBER
mkraft@brown.edu

Matthew G. Springer
University of North Carolina - Chapel Hill
mgspringer@unc.edu

I. Introduction

The returns to improved performance evaluation systems have long been of interest to economists and employers. Evaluation systems have the potential to better align workers' effort with organizational goals as well as to inform employee skill development (Gibbons 1998; Prendergast 1999; Oyer and Schaefer 2011). We study efforts to strengthen performance evaluation in the K-12 public education system, one of the largest economic sectors in the U.S. employing more than 3.5 million teachers. 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 2022). 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.

In this paper, we examine how the statewide implementation of newly reformed teacher evaluation systems affected student achievement and educational attainment. Between 2009 and 2017, 44 states and Washington, DC implemented major reforms to their teacher evaluation systems. Prior to the reforms, teacher evaluation was an infrequent and largely perfunctory exercise that resulted in nearly all teachers receiving satisfactory ratings (Weisberg et al. 2009). Strong incentives by the federal government spurred states to reform evaluation systems by adding regular teacher evaluations based on multiple measures (including student achievement growth) and using performance ratings to inform professional development and personnel decisions.

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. Our 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. Our primary estimates capture the *average* effect of these reforms on a national scale, pooling across different system designs and implementation approaches. We further explore potential heterogeneity in these effects based on the evaluation metrics and design features adopted by states.

Understanding the *average* effect of state teacher evaluation reform efforts at the national level is critical from a policy perspective. High-stakes teacher evaluations were a central education priority of the Obama administration and one of the signature federal education reform efforts of the last several decades (McGuinn 2016). These reforms were also highly controversial, leading to protests and lawsuits challenging their legitimacy in several states (McGuinn 2012; Government Accountability Office 2015a; Sawchuk 2015; Paige 2020). Efforts to implement teacher evaluation reforms came with substantial financial and time costs as well. A back-of-the-envelope estimate suggests that public schools spend about $2 billion each year on teacher evaluation systems.[1] New evaluation systems also created large demands on administrators' time to conduct frequent observations and complete considerable paperwork (Neumerski et al. 2018; Kraft and Christian 2022), amounting to as much as 19 total days of work per school year (Hess and Bell 2017).

[1] Chambers and colleagues (2013) estimate that the costs of implementing teacher evaluation systems in three large school districts was about four tenths of a percent of their total expenditures. Four thenths of total public school expenditures in the U.S. ($604 billion in 2011 and $601 billion in 2012) is approximately $2.4 billion.

Our study also provides a powerful window into the "scale-up problem" in the context of the decentralized U.S. public education system (List 2022). Several studies of large urban districts in the vanguard of the reform effort provide evidence that rigorous teacher evaluation systems can improve teacher performance and student achievement (Taylor and Tyler 2012; Dee and Wyckoff 2015; Steinberg and Sartain 2015; Adnot et al. 2016; Sartain and Steinberg 2016; Dotter, Chaplin, and Bartlett 2021; Dee, James, and Wyckoff 2021). Personnel management practices such as measuring performance, dismissing low performing teachers, and rewarding high performers have also been shown to positively impact student achievement across schools in large international samples (Bloom et al. 2015; Lemos, Muralidharan, and Scur 2021). At the same time, prior research has found that intensive "best-practices" management interventions have failed to produce or sustain gains in student achievement when taken to scale (Fryer 2017; Garet et al. 2018; Stecher et al. 2018; Muralidharan and Singh 2020).

We find that, on average, state teacher evaluation reforms had no discernible 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. These estimates are precise enough to rule out the predicted positive effects of teacher evaluation reforms on student achievement that come from simulation and dynamic models in which 10 to 20 percent of first- or second-year teachers are dismissed (roughly 0.5 to 1 percent of all K-12 public school teachers) based on value-added scores (Goldhaber and Hansen 2010; Staiger and Rockoff 2010; Rothstein 2015; Liebowitz 2021).[2] We also find no evidence

[2] Goldhaber and Hansen (2010) find that dismissing the bottom 25 percent of first- and second-year teachers in the distribution of value-added scores would raise student achievement by 0.025 SD. Staiger and Rockoff (2010) find

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 examine the robustness of these null results in several ways. First, we replicate our null findings using newly developed two-way fixed effects (TWFE) estimators that address potential bias in the presence of heterogeneous treatment effects and staggered treatment adoption (Callaway and Sant'Anna 2020; Goodman-Bacon 2021; Sun and Abraham 2021). Second, we address the potential conflation of evaluation reforms with other related efforts to increase teacher accountability and a wide range of time-varying education reforms that occurred during our panel period. Our results are essentially unchanged when we control directly for these other state-level policy reforms.

We then focus on exploring whether these average estimates mask important treatment effect heterogeneity based on variation in evaluation system designs across states. Specifically, we construct a state-level index of evaluation system design rigor based on 10 evaluation design elements commonly identified as key features of effective systems (Doherty and Jacobs 2015; Howell and Magazinnik 2017).[3] We also group system design elements into three broad categories motivated by the primary mechanisms through which proponents argued evaluation would benefit students. Overall, we find little evidence of heterogeneity based on either our index approach or using the 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.

---

that dismissing 40 percent of first-year teachers would raise student achievement by 0.045 SD. Rothstein (2015) finds that dismissing 20 percent of second-year teachers would increase student achievement by 0.018 SD. These findings suggest an approximate linear relationships where dismissing 10 to 20 percent of pre-tenure teachers would raise student achievement by 0.01 to 0.02 SD. All three studies assume dismissed teachers would be replaced with early-career teachers of average ability.

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

What might account for these null findings? We draw on research examining teacher evaluation reforms through the lenses of political science, organizational theory, and the science of scaling to unpack the failure of evaluation reforms at the national level. These literatures identify four key explanations: political opposition, the decentralized structure of K-12 public education in the U.S., capacity constraints, and limited generalizability. Economic theory also suggests that effective evaluation reforms require a grand bargain with teachers: more accountability in exchange for greater compensation (Goldhaber and Hansen 2010). We draw on nationally representative survey data to show that evaluation reforms decreased teachers' job satisfaction, autonomy, and perceived job security without offsetting these non-pecuniary costs with increased pay.

Our paper makes three primary contributions to the literature. First, and foremost, our nationally representative study provides the broadest and most generalizable evidence on the efficacy of teacher evaluation reforms in the U.S. and among the first evidence on how these reforms affected students' long-term outcomes. Existing evidence on the effects of teacher evaluation reforms is mixed and limited to a narrow set of districts/states and short-term outcomes (Taylor and Tyler 2012; Dee and Wyckoff 2015; Steinberg and Sartain 2015; Adnot et al. 2016; Sartain and Steinberg 2016; Dotter, Chaplin, and Bartlett 2021; Dee, James, and Wyckoff 2021; Taylor 2022). Examining longer-term outcomes of education interventions is critical given that such effects can appear even when effects on test scores are not present or fade out (Chetty et al. 2011; Bailey et al. 2020). Our well-identified and well-powered null effects provide valuable information in a setting where prior evidence has documented positive effects on a smaller scale (Abadie 2020) and where these results have catalyzed substantial political, financial, and organizational investments in reforming teacher evaluation systems.

Second, we contribute to the broader economics literature on personnel management practices (e.g., Bloom et al. 2015) by providing new evidence on the impact of evaluation systems on worker productivity in the public sector (Baker 1992; Gibbons 1998; Bloom and Van Reenen 2007, 2011; Heinrich and Marschke 2010; Heinrich, Meyer, and Whitten 2010; Cappelli and Conyon 2018).

Third, we contribute to the cross-disciplinary literature on the science of scaling evidence-based interventions (Coburn 2003; Honig 2006; Manna 2010; Banerjee et al. 2017; Davis et al. 2017; Muralidharan and Niehaus 2017; Al-Ubaydli, List, and Suskind 2020; Gupta et al. 2021; Zhou et al. 2021). Our work is most closely related to Jepsen and Rivkin's (2009) seminal study of how supply-side constraints made efforts to reduce class sizes at scale across California largely ineffectual. High-stakes teacher evaluation systems went from being a policy proposal piloted in DC public schools to a nationwide reform initiative with unprecedented speed and federal support. We replicate and extend prior research that teacher evaluation systems can be effective by documenting how a select group of states and districts identified ex-post as exemplary did appear to raise student achievement. We also show how, at scale, even state reforms that shared similar design features as these exemplar systems failed to impact student outcomes. Our findings illustrate an underdiscussed dimension of the scale-up problem, the tradeoff between the rapid implementation of top-down reform efforts and the political, organizational, and logistical challenges this approach creates at the local level.

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, robustness checks in VI, explore potential explanations for our null findigns in VII, and conclude in Section VIII with a discussion of the implications of our results for policy and practice.

II. Background

Conceptual Framework

The widespread adoption of teacher evaluation reforms marked a shift from evaluation systems that relied primarily on perfunctory 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 $4.35 billion federal Race to the Top (RTTT) program and its offer of large competitive grants to states that were struggling during the Great Recession (Howell and Magazinnik 2017). The rubric for evaluating RTTT applications rewarded states for using student outcomes to evaluate teachers and inform personnel decisions with evaluation ratings. The Obama administration also required states to commit to implementing teacher evaluation reforms in exchange for a waiver from the No Child Left Behind (NCLB) mandate to reach 100% proficiency by 2014.

Reformers envision two primary mechinisms for how new teacher evaluation systems would improve instruction and achievement. First, evaluation reforms have the potential to change the composition of the teacher workforce by tying high-stakes personnel decisions such as dismissal and tenure decisions to performance ratings (Gordon, Kane, and Staiger 2006; Goldhaber and Hansen 2010; Staiger and Rockoff 2010; Rodriguez, Swain, and Springer 2020; Liebowitz 2021; Sartain and Steinberg 2021). Second, teacher evaluation may directly improve current teacher 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 (Firestone 2014; Donaldson and Papay 2015). 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 (Mintrop and Trujillo 2007; Springer 2010; Woulfin and Rigby 2017; Donaldson 2020; Donaldson and Firestone 2021; Galey-Horn and Woulfin 2021).

Underlying these mechnisms is a multi-stage policy diffusion process for how the federal government aimed to influence the composition and practices of the teacher workforce across the decentralized U.S. education system. As McGuinn (2012) describes, the Obama administration pivoted away from the focus on sanctions to compel action under No Child Left Behind toward incentives to spur state policy change. This state policy change was intended to lead to direct changes in district policies that are enacted in ways to produce behavioral and personnel changes among school leaders and teachers. Finally, these behavioral and personnel changes would then result in improved instructional quality and ultimately greater student achievement.

Research on Implementation

Numerous studies confirm that the federal government successfully leveraged the RTTT grant competition and NCLB waivers to spur widespread changes to state laws and teacher evaluation policies (McGuinn 2012; Wong 2015; NCTQ 2016; Howell and Magazinnik 2017, 2020; Bleiberg and Harbatkin 2020). A study of the first 19 states to win the RTTT grant competition found that the vast majority were induced to 1) assign weights to student achievement growth in evaluations, 2) adopt multicategory rating systems, 3) conduct annual evaluations, 4) require evaluations to be used for professional development, and 5) require

evaluations to be used for dismissal decisions (Dragoset et al. 2016). Far fewer states required evaluations to be used for compensation and career advancement decisions.

Evidence also suggests that state evaluation reforms meaningfully impacted evaluation practices on the ground. For example, survey data collected by the U.S. Government Accountability Office suggests that most district leaders felt the reforms introduced under RTTT were implemented with moderate (36%) or high quality (40%) (Government Accountability Office 2015b). Data from the National Council for Teacher Quality (NCTQ) teacher contract database collected in 2019 provides additional evidence that many large districts implemented evaluation reforms in alignment with federal policy guidance (2022). In particular, the vast majority of large districts required teachers to receive feedback on evaluations (91%) and required annual evaluations for non-tenured teachers (86%). Most (75%) assigned at least some weight to student achievement—a heavily weighted feature on the RTTT application scoring rubric—and used teacher evaluation as a criteria for dismissal (61%), and to a lesser degree, offered bonuses (42%) for strong evaluations.

At the same time, certain features of the new high-stakes evaluation systems promoted by the federal government were taken up less often than others. For example, nationally, fewer than one percent of teachers were rated as unsatisfactory under the new evaluation systems, with formal performance-based dismissals being 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 low enough that most teachers qualified for the bonuses (NCTQ 2019). These challenges arose, in part, because reformers failed to engage key stakeholders at times. Roughly one-third of state RTTT winners reported challenges maintaining support for the reforms from state legislatures and teachers' unions (Government Accountability Office 2015b).

While research suggests many states and districts did engage in meaningful efforts to reform their teacher evaluation practices on the ground (Howell and Magazinnik 2017), the degree to which different design features were implemented also varied considerably (Kraft and Gilmour 2017).

Overall, state reforms led districts to implement new teacher evaluation systems where teachers are evaluated annually on a multi-category scale based on administrators' ratings on an instructionally aligned observation rubric and, in most cases, measures of student growth. Administrators typically provide some individualized performance feedback (often written) to teachers and use evaluation ratings to inform professional development and dismissal decisions. Few teachers are actually removed for poor performance, but teachers generally perceive dismissal as a threat and those rated below satisfactory leave at higher rates. While some of these evaluation features existed in pockets prior to the state reforms (Hallgren, James-Burdumy, and Perez-Johnson 2014), their prevalence in districts nationwide suggests a marked shift in teacher evaluation policy.

Evidence on Teacher Evaluation and Student Achievement

Existing empirical research on teacher evaluation provides insights into the mechanisms through which effective personnel evaluation systems can raise student achievement. Several studies in large urban school districts point to the potential for evaluation systems to serve as engines for professional growth. Experimental studies of low-stakes observation and feedback by peers (Papay et al. 2020; Burgess, Rawal, and Taylor 2021) and administrators (Garet et al. 2018) have found positive effects on achievement. Cincinnati Public Schools' peer evaluation and feedback system in which teachers were observed and evaluated by experienced, expert teachers and school principals improved teachers' ability to raise student achievement in math

but not ELA (Taylor and Tyler 2012). 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 2021). However, field trials of training programs designed to improve evaluator feedback in high-stakes settings found no improvements on feedback quality or student achievement (Mihaly et al. 2018; Kraft and Christian 2022), 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).

Other studies document how high-stakes evaluation can lead to positive changes in the composition of the teacher workforce. Several studies have found that new teacher evaluation systems increased voluntary turnover among lower-performing teachers (Loeb, Miller, and Wyckoff 2015; Steinberg and Sartain 2015; Rodriguez, Swain, and Springer 2020; Cullen, Koedel, and Parsons 2021). Similarly, evidence from a national study of teacher evaluation reforms found that these reforms increased the number of new teacher candidates who had attended more competitive undergraduate institutions but also decreased the overall supply of teacher candidates (Kraft et al. 2020).

Perhaps the most compelling evidence for the efficacy of high-stakes teacher evaluation comes from the District of Columbia Public Schools' DC IMPACT program. 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; Dotter, Chaplin, and Bartlett 2021; Dee, James, and Wyckoff 2021).

Evidence from studies examining teacher evaluation systems that are more representative of those adopted at scale in the U.S. is decidedly more mixed. In an experimental study of the new teacher evaluation system in Chicago Public Schools, Steinberg & Sartain (2015) found the pilot program 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 2021).

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,

DC was the first to reform its evaluation system, in 2009, while states implemented reforms to their teacher evaluation systems between 2012 to 2017 (See Appendix Figure A1).[4] 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.

We also collected data on 10 teacher evaluation design features identified in the literature as key features of evaluation systems (NCTQ 2011, 2019; Doherty and Jacobs 2015; Howell and Magazinnik 2017). We then constructed an index equal to the number of teacher evaluation policy design features that states required districts to implement (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 design features, we group the 10 design features into three categories based on their policy rationales: improving the measurement of teacher performance; using performance measures for accountability and incentives; and using performance measures to provide feedback and inform professional development (See Appendix Table B3). We then create non-mutually exclusive indicators for whether each state implemented a majority of the design features in a given category. Sixteen states adopted a majority of design features focused on enhancing the reliability of teacher evaluation measures, 19 adopted design features focused on using performance measures for accountability and incentives, and 29 adopted design features focused on using evaluations to provide feedback and inform professional development.

[4] Washington, DC does not contribute to the estimated effect of teacher evaluation on achievement because we do not observe pre-treatment math or ELA scores. We run a parallel set of models using NAEP that do include Washington, DC and find similar results. We also observe pre-treatment attainment outcomes and leverage data from Washington, DC to identify those effects.

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).[5] 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.[6] Table 1 displays descriptive statistics for the full sample. We observe about 460,000 district-grade observations in math and 490,000 district-grade observations in ELA.

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 percent of 18-year-olds who earned a high school diploma (including a GED) in a given year by state of birth and apply appropriate PUMS person weights. To measure college enrollment for each state and year, we calculate the percent of 22-year-olds who ever enrolled in a college in a given year by state of birth, again using PUMS person weights from 2008 to 2020.[7] This procedure follows recent research on state education reforms that measures

[5] 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.
[6] 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. This includes public charter schools but excludes private schools.
[7] In Appendix Figure A2, we confirm that our results are unchanged when we use the percent of high school graduates by age 20 (to account for students who repeat grades or do not graduate with their ninth-grade cohort) and the percent of college graduates by age 20 (to increase the number of relative time periods post-treatment in which cohorts were exposed to K-12 evaluation reforms).

educational attainment based on the expected degree-earning age (Murnane 2013; Jackson, Wigger, and Xiong 2021; Rothstein and Schanzenbach 2021).[8]

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. 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

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 function 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 linear 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 (Wooldridge 2021). It also improves the precision with which we can identify null effects. 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 function covariates include county level GDP, a poverty index, county

[8] 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. A strength of using the ACS to measure attainment is that the same protocol is used to collect the data across time. One limitation is that a single member of the household reports educational attainment data for every member of the household.

unemployment rate, district-level student-teacher ratio, and district-level per-pupil expenditures.[9] We also add covariates for baseline outcomes interacted with linear time trends to control for dynamic differences in student achievement and attainment based on pre-treatment outcomes. 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. Methods

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:

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

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]. We set this term to zero for all states that never implemented evaluation reforms. $\boldsymbol{X}$ is a vector of baseline covariates including the school district characteristics, education production function 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

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

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.[10]

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.[11]

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 replace our district level controls with 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.

[10] An alternative approach to estimating standard errors using the wild cluster bootstrap, which accounts for the small number of state clusters, produces very similar confidence intervals in our setting (Cameron, Gelbach, and Miller 2008; Roodman et al. 2019).

[11] Appendix Table A1 describes the number of treated states and observations across relative time. The analytic sample is "trimmed" to mitigate weak panel balance. As shown in Appendix Figure A3 results remain consistent when we estimate event study models with the untrimmed sample that includes distal pre and post estimates.

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) comparison states provide a valid counterfactual for the trends in treated states in the absence of treatment; and 2) there are no unobserved factors correlated with both our outcomes of interest and the timing of state-wide implementation of teacher evaluation reforms across states. We examine the first assumption visually and empirically using the non-parametric event study and also by estimating a separate DiD model that includes state-specific linear time trends. We 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.[12]

Several recent studies have shown that estimates from standard event study 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 2021). Consequently, we also report results from alternative TWFE estimators robust to issues related to heterogeneous treatment effects (Cengiz et al. 2019; Baker,

[12] In auxiliary DiD models, we also add frequency weights for student enrollment. The weighted models yield similarly sized null effects.

Larcker, and Wang 2021; Sun and Abraham 2021). 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 student achievement 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.[13] 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 small in magnitude and 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. Furthermore, as shown in Appendix Table A4, our DiD estimates are robust to alternative specification of control vectors including district time-varying controls (Panel A) time-varying state-level controls (Panels B and C).

[13] The event study estimates with and without controls yield similar estimates (see Appendix Table A2 and Table A3).

Educational Attainment

Event study and DiD estimates similarly 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, both of which are small in magnitude and indistinguishable from zero. Importantly, we once again find no evidence of differential pre-treatment trends. Estimates from event study models are precise enough to rule out a 2.5 percentage point increase in high school graduation and college enrollment across all observed years post-reform (See Appendix Table A4).

The event study estimates map out the cumulative effect of exposure to evaluation reforms during a students' K-12 education. We might expect to find effects emerging for only those cohorts that could have benefitted from new teacher evaluation systems for multiple years. We observe as many as seven total years of exposure to K-12 teacher evaluation reforms for high school graduation outcomes at age 18 and three total years of exposure for college enrollment outcomes at age 22 (estimates from τ_0 to τ_3 reflect cohorts that graduated from high school before the reforms were implemented). However, we find no evidence that students exposed to evaluation reforms for longer periods experienced larger effects. If anything, estimates trend in a slight negative direction for those cohorts exposed to evaluation reforms for longer periods of time. As shown in Appendix Figures A2 and A3, when we examine college enrollment effects among cohorts with even longer exposure to K-12 evaluation reforms by using 20 year olds or expanding our event study window, estimates remain small and insignificant.

DiD results also show a null effect of teacher evaluation on educational attainment. Table 2, Panel B, presents the DiD estimates for educational attainment, pooling the treatment effect estimate over all post-treatment years. Our most precise estimates from models with covariates

allow us to rule out effects as small as a 1 percentage point increase in high school graduation and college enrollment.[14] As shown in Appendix Table A4 Panel B, similar to our achievement results, we also find that our null effects are robust to the inclusion of alternative control sets using time-varying state-level covariates.

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.[15] Overall, we find no evidence that high design rigor evaluation systems positively affected student achievement or educational 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 implemented teacher evaluation reform but required districts to put very few design features 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].[16]

[14] Appendix Table A5 present the DiD estimates of the effect of teacher evaluation on college completion measured at age 24 and once again all results remain consistent.
[15] 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 features we observe in our index.
[16] 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)).

The results in Table 3 suggest that, in general, the effect of teacher evaluation reforms on student outcomes did not vary by the rigor of teacher evaluation system designs.[17] We provide further evidence of these null effects by plotting event studies for states with a high number of design features compared with states with a low number of design features 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 designs (i.e., between one and six teacher evaluation design features) 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 some evidence of small decreases in ELA scores one to two years after treatment for states with weak evaluation designs. As shown in Appendix Table A6, 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.[18]

Next, we use the 10 design features 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 for states that adopted design

[17] The results in Table 3 are similar when we use the first principal component from Principal Components Analysis.
[18] As shown in Appendix Figure A4, we find similarly precise null effects when we change the definition of high quality to states that implemented eight or more teacher evaluation design features.

features to improve the reliability of teacher evaluation measures (e.g., use student test scores, 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, grant tenure). Figure 4, Panel C displays estimates for states that used teacher evaluation to inform professional development or provide 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 provide little evidence that the effect of teacher evaluation reform varied with specific design features; the estimated coefficients are generally small in magnitude and statistically insignificant.

Finally, in Figure 5, we attempt to reconcile our consistent null results with prior research documenting the positive effects of evaluation reforms in selected districts. In October of 2018, the National Council for Teacher Quality (NCTQ) released a report that profiled two states and four school districts that were judged to have designed and implemented exemplary evaluation systems. 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). According to NCTQ, 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 exemplar systems by fitting models where we include two mutually exclusive indicators identifying 1) states and districts that were identified as having exemplar systems and 2) all other states that adopted reforms. 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 a high of 0.15 SD. In our pooled DiD model, we estimate a marginally significant positive effect of 0.09 SD in math and 0.07 SD effect in ELA (See Appendix Table A7).

An important caveat to these analyses is that exemplar districts were selected ex-post by NCTQ based partly on their outcomes. Consequently, the results presented in Figure 5 and Appendix Table A8 are best understood as descriptive rather than causal. Nevertheless, we view these results as providing evidence that is consistent with prior research. We find positive impacts of teacher evaluation reform in a small number of select states and districts, while also detecting null effects among the vast majority of states and districts that implemented reforms that were more representative of those adopted at scale nationally.

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 A9, 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 free or reduced price lunch (FRPL), percent Black, and percent 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 with two exceptions. Specifically, in Panel B, the estimated coefficient on the percent Hispanic interaction for ELA is negative and statistically significant and in Panel C, the estimated coefficient on the FRPL interaction for high school graduation is negative and marginally significant. This implies that, if anything, the reforms may have widened rather than closed achievement gaps, although the size of the effect is substantively small. The results in Appendix Table A9 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 1 SD increase (20 percentage points) in the percent of FRPL students results in a 1.7 percentage point decrease in high school graduation.

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. Our first approach utilizes a stacked DiD estimator that is robust to heterogeneous treatment effects in models with staggered timing of adoption (Cengiz et al. 2019; Baker, Larcker, and Wang 2021). Specifically, we create six datasets, one for each cohort of states that reformed their teacher evaluation systems in the same year (i.e., 2012, 2013, 2014, 2015, 2016, 2017), respectively, 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 and multi-way cluster our standard errors by state and cohort. Our second approach estimates cohort-specific average treatment effects on the treated (CATT) using the estimator developed by Sun and Abraham (2021). 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 that compare late to early reformers and over-weighting the effects of early adopters relative to later adopters.

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 and our main analytic strategy. 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.

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 A10 includes results for the achievement and attainment outcomes with and without covariates augmented with state-specific linear trends.[19] The achievement results are within 0.01 SD of the main results in Table 2. Similarly, the attainment results differ by less than 1 percentage point from the main results.

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

[19] 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.

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).[20]

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 include a vector of 19 time-varying education policies (Howell and Magazinnik 2017; Kraft et al. 2020). As shown in Appendix Table A11, 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 attainment outcomes. The precisely estimated null effects suggest that contemporaneously adopted reforms do not appear to bias our estimated effects of teacher evaluation.

Replicating Results in NAEP

We use SEDA data 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 (Booher-Jennings 2005; Springer 2008; Neal and Schanzenbach 2010; Ballou and Springer 2017). 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 A12, consistent with our main results, we find null effects on achievement. We can rule out positive

[20] See Kraft et al. (2020) for a complete listing of the education policy reforms that occurred contemporaneously during the sample timeframe.

effects as small as 0.01 SD in math and 0.02 SD in ELA in models including controls.[21] These results add further support for our primary analyses using the SEDA.

VII. Potential Explanations for Null Effects

Consistent with the results of prior studies, we find positive effects of similar magnitude on student achievement for a small set of states and districts with systems identified as exemplary ex-post (Taylor and Tyler 2012; Dee and Wyckoff 2015; Adnot et al. 2016; James and Wyckoff 2020; Dotter, Chaplin, and Bartlett 2021). This leads naturally to the question of why, at the national level, teacher evaluation reforms appear to have had little impact on student achievement or educational attainment. We examine this question by drawing on research from political science, organizational theory, and the science of scaling as well as by conducting a range of exploratory analyses using nationally representative data on public school teachers' experiences on the job and compensation. We highlight five primary factors that likely undercut the efficacy of teacher evaluation reforms at a national scale: political opposition, the structure of public education, capacity constraints, limited generalizability, and sanctions without increased compensation.

Political Opposition

The teacher evaluation reforms advanced by the Obama administration were controversial. While the strong financial incentives of the RTTT grant competition were enough to compel many states to make meaningful changes to state laws and policy, states that were not actively supportive of these changes engaged in bad-faith efforts of symbolic compliance to

[21] These models control for the same baseline district characteristics interacted with linear time trends 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 interacted with linear time trends, and an indicator for whether a school made Adequate Yearly Progress in 2003 (Reback et al. 2013).

compete for funding with minimal changes on the ground (McGuinn 2012). The top-down push for rapid change also failed to involve local governments and teachers' unions, two critical stakeholders. Ultimately, teachers' unions challenged the legitmacy of the new systems in more than a dozen states and succeeded in limiting their implementation (Paige 2020). Many states also struggled to maintain support among governors, state boards of education, and local legislatures as the political cycle led to frequent churning of state leaders.

The Decentralized K-12 Public Education System

The nature of the decentralized U.S. K-12 public education system also constrains what the federal government can accomplish via top-down reforms. The limited influence of an incentive-based approach to compelling states to act meant that states were afforded considerable flexibility, resulting in wide variability in specific design features across states and differing degrees of implementation (Howell and Magazinnik 2017). Reformers failed to distinguish between foundational design elements that were non-negotiable and those that could be adapted to the local context. The loose coupling of education policy and practice further undercut good-faith efforts by states that were pre-disposed to favor or had taken steps to develop high-stakes evaluation systems. The effectiveness of teacher evaluation was also dependent on how administrators in each state chose to interpret and enact these policies in schools.

Capacity Constraints

Evidence suggests that capacity constraints at the local level played an important role in limiting the effectiveness of the reforms (Murphy, Hallinger, and Heck 2013). Despite large initial investments by the federal government, district leaders reported that financial capacity constraints (e.g., human resource, capacity) created implementation challenges and about one-third of districts modified their original plans as a result of these challenges (Government

Accountability Office 2015b). One concrete consequence of financial constraints was that districts overwhelmingly tasked principals with the responsibility of observing, evaluating, and providing feedback to teachers rather than hiring additional staff to serve as master evaluators (Herlihy et al. 2014). Many principals were ill-prepared to provide substantive feedback to teachers working across a range of grades and subjects, which led to a narrowing of the scope, depth, and quality of feedback teachers received (Kraft and Christian 2022; Hunter and Springer 2022). Administrators also had limited capacity to conduct frequent observations and feedback meetings. Adminstrators prioritized the considerable paperwork the new systems required because it was easier for districts to monitor (i.e., bureaucratic compliance) (Neumerski et al. 2018). This focus on reporting requirements and paperwork is consistent with previous research on large-scale managerial reforms in public education (Muralidharan and Singh 2020).

Limited Generalizability

Another possible explanation for the failure of evaluation reforms at the national level is the lack of generalizability underlying both the theory behind teacher evaluation and early evaluations of teacher evaluation reforms. A central assumption of high-stakes teacher evaluation systems is that there is an elastic supply of average-ability novice teachers who can quickly fill open vacancies. However, teacher labor supply varies dramatically across local markets making this assumption unrealistic in many settings (Edwards et al. 2022). Early efforts to implement these reforms in places like Washington DC and New York City, where districts were under mayoral control, were also unrepresentative of the state-wide contexts in which the reforms played out.

Sanctions without Increased Compensation

Economic theory suggests evaluation reform would only affect average teacher quality if salaries were increased enough to offset any increased risk associated with becoming a teacher (Goldhaber and Hansen 2010). Prior research documents that teacher accountability reforms decreased job satisfaction and perceived autonomy among new teachers (Kraft et al. 2020). We extend this work in Table 4 by examining how teacher evaluation reforms affected satisfaction and autonomy across all public school teachers using nationally representative survey data collected as part of the Schools and Staffing Survey/National Teacher and Principal Survey. Using the same DiD framework as in our student achievement and educational attainment models, we find that teacher evaluation reforms decreased the proportion of teachers who strongly agreed that they were satisfied with their job and who felt their jobs were secure. Teacher evaluation also significantly decreased the proportion of teachers who were satisfied with their pay. We find no evidence that teacher evaluation affected the control teachers felt they had over instructional content, material, and techniques. These finding suggests that in addition to reducing teacher job satisfaction, the reforms did not appear to offset negative effects by meaningfully changing important elements of teacher practice—a key mechanism in the theory of change for teacher evaluation reform. We also find no evidence that teachers' actual compensation increased in parallel with the implementation of teacher evaluation (see Table A13). Taken together, these results suggest that teacher evaluation reforms had unintended negative consequences for teacher experiences on the job that were not offset by any increases in compensation or changes in teacher practice.

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 the 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.

We draw on research and theory from economics, political science, organizational theory, and the science of scaling to understand the likely reasons why evaluation reforms failed to improve student achievement or educational attainment at the national level. Based on that literature, we identify five likely reasons why evaluation reforms that were shown to be successful in a handful of districts such as Washington, DC, were unable to be replicated at scale, namely: political opposition, the decentralized structure of K-12 public education in the U.S., capacity constraints, limited generalizability, and the lack of increased teacher compensation to offset the non-pecuniary costs associated with high-stakes teacher evaluations.

Adopting evaluation systems like the one implemented in Washington DC requires a significant investment of time, money, human and political capital. Many states and districts may have believed that the costs of fully adopting high-stakes evaluations outweighed the benefits, and instead did so in more superficial ways. Other states and districts that were invested in the new systems may have lacked the internal capacity to implement and sustain the reforms. Ultimately, high-stakes teacher evaluation reforms appear to have been organizationally, economically, and politically challenging to scale with fidelity in the absence of local actors to champion the reforms from the ground-up. This was rarely the case, resulting in a large financial

investment in a reform that had little effect on student achievement or educational attainment at the national scale.

Reference List

Abadie, Alberto, "Statistical nonsignificance in empirical economics," *American Economic Review: Insights*, 2 (2020), 193–208.

Adnot, Melinda, Dee, Thomas S., Katz, Veronica, and Wyckoff, James, "Teacher turnover, teacher quality, and student achievement in DCPS," *Educational Evaluation and Policy Analysis*, 39 (2016), 54–76.

Al-Ubaydli, Omar, List, John A., and Suskind, Dana, "2017 Klein Lecture: The Science Of Using Science: Toward An Understanding Of The Threats To Scalability," *International Economic Review*, 61 (2020), 1387–1409 (Wiley Online Library).

Anderson, Kaitlin P., Cowen, Joshua M., and Strunk, Katharine O., "The impact of teacher labor market reforms on student achievement: Evidence from Michigan," *Education Finance and Policy*, 1 (2021), 1–43.

Bailey, Drew H., Duncan, Greg J., Cunha, Flávio, Foorman, Barbara R., and Yeager, David S., "Persistence and fade-out of educational-intervention effects: Mechanisms and potential solutions," *Psychological Science in the Public Interest*, 21 (2020), 55–97 (Sage Publications Sage CA: Los Angeles, CA).

Baker, Andrew, Larcker, David F., and Wang, Charles C.Y., "How Much Should We Trust Staggered Difference-In-Differences Estimates?," SSRN (2021).

Ballou, Dale, and Springer, Matthew G., "Has NCLB encouraged educational triage? Accountability and the distribution of achievement gains," *Education Finance and Policy*, 12 (2017), 77–106.

Banerjee, Abhijit, Banerji, Rukmini, Berry, James, Duflo, Esther, Kannan, Harini, Mukerji, Shobhini, Shotland, Marc, and Walton, Michael, "From proof of concept to scalable policies: Challenges and solutions, with an application," *Journal of Economic Perspectives*, 31 (2017), 73–102.

Bleiberg, Joshua, and Harbatkin, Erica, "Teacher Evaluation Reform: A Convergence of Federal and Local Forces," *Educational Policy*, 34 (2020), 918–952.

Bloom, Nicholas, Liang, James, Roberts, John, and Ying, Zhichun Jenny, "Does Working from Home Work? Evidence from a Chinese Experiment *," *The Quarterly Journal of Economics*, 130 (2015), 165–218.

Booher-Jennings, Jennifer, "Below the bubble:'Educational triage' and the Texas accountability system," *American Educational Research Journal*, 42 (2005), 231–268.

Briole, Simon, and Maurin, Éric, "There's always room for improvement: the persistent benefits of repeated teacher evaluations," (Padua, Italy, 2021).

Burgess, Simon, Rawal, Shenila, and Taylor, Eric S., "Teacher Peer Observation and Student Test Scores: Evidence from a Field Experiment in English Secondary Schools," *Journal of Labor Economics*, 39 (2021), 1155–1186.

Callaway, Brantly, and Sant'Anna, Pedro HC, "Difference-in-differences with multiple time periods," *Journal of Econometrics*, (2020).

Cengiz, Doruk, Dube, Arindrajit, Lindner, Attila, and Zipperer, Ben, "The effect of minimum wages on low-wage jobs," *The Quarterly Journal of Economics*, 134 (2019), 1405–1454.

Chambers, J., Brodziak de los Reyes, I., and O'Neil, C., "How much are districts spending to implement teacher evaluation systems," (Washington, D.C., RAND, 2013).

Chetty, Raj, Friedman, John N., Hilger, Nathaniel, Saez, Emmanuel, Schanzenbach, Diane Whitmore, and Yagan, Danny, "How does your kindergarten classroom affect your

earnings? Evidence from Project STAR," *The Quarterly journal of economics*, 126 (2011), 1593–1660 (MIT Press).

Chetty, Raj, Friedman, John N., and Rockoff, Jonah E., "Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates," *The American Economic Review*, 104 (2014), 2593–2632.

Coburn, Cynthia E., "Rethinking scale: Moving beyond numbers to deep and lasting change," *Educational Researcher*, 32 (2003), 3–12.

Cullen, Julie Berry, Koedel, Cory, and Parsons, Eric, "The compositional effect of rigorous teacher evaluation on workforce quality," *Education Finance and Policy*, 16 (2021), 7–41.

Davis, Jonathan MV, Guryan, Jonathan, Hallberg, Kelly, and Ludwig, Jens, "The economics of scale-up," (National Bureau of Economic Research, 2017).

Dee, Thomas S., James, Jessalynn, and Wyckoff, Jim, "Is Effective Teacher Evaluation Sustainable? Evidence from District of Columbia Public Schools," *Education Finance and Policy*, 16 (2021), 313–346.

Dee, Thomas S., and Wyckoff, James, "Incentives, selection, and teacher performance: Evidence from IMPACT," *Journal of Policy Analysis and Management*, 34 (2015), 267–297.

Doherty, Kathryn, and Jacobs, Sandi, "State of the States 2015: Evaluating Teaching, Leading and Learning," (Washington, D.C., National Council on Teacher Quality, 2015).

Donaldson, Morgaen, "Multidisciplinary Perspectives on Teacher Evaluation: Understanding the Research and Theory," (New York, Routledge, 2020).

Donaldson, Morgaen, and Firestone, William, "Rethinking teacher evaluation using human, social, and material capital," *Journal of Educational Change*, 22 (2021), 1–34.

Donaldson, Morgaen, and Papay, John, "Teacher Evaluation for Accountability and Development," in *Handbook of Research in Education Finance and Policy* (Routledge, 2015).

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

Dragoset, Lisa, Thomas, Jaime, Herrmann, Mariesa, Deke, John, James-Burdumy, Susanne, Graczewski, Cheryl, Boyle, Andrea, Tanenbaum, Courtney, Giffin, Jessica, and Upton, Rachel, "Race to the Top: Implementation and Relationship to Student Outcomes," *National Center for Education Evaluation and Regional Assistance*, (2016) (ERIC).

Edwards, Danielle Sanderson, Kraft, Matthew A., Christian, Alvin, and Candelaria, Christopher A., "Teacher Shortages: A Unifying Framework for Understanding and Predicting Vacancies," *EdWorkingPapers.com* (Annenberg Institute at Brown University, 2022).

Firestone, William A., "Teacher evaluation policy and conflicting theories of motivation," *Educational Researcher*, 43 (2014), 100–107.

Fryer, Roland G., "Management and student achievement: Evidence from a randomized field experiment," (National Bureau of Economic Research, 2017).

Galey-Horn, Sarah, and Woulfin, Sarah l, "Muddy Waters: The Micropolitics of Instructional Coaches' Work in Evaluation," *American Journal of Education*, 127 (2021), 441–470.

Garet, Michael S., Wayne, Andrew J., Brown, Seth, Rickles, Jordan, Song, Mengli, and Manzeske, David, "The Impact of Providing Performance Feedback to Teachers and Principals," (Washington, D.C., National Center for Education Evaluation and Regional Assistance, 2018).

Gibbons, Robert, "Incentives in organizations," *Journal of Economic Perspectives*, 12 (1998), 115–132.

Goldhaber, Dan, and Hansen, Michael, "Using performance on the job to inform teacher tenure decisions," *American Economic Review*, 100 (2010), 250–55.

Goodman-Bacon, Andrew, "Difference-in-differences with variation in treatment timing," *Journal of Econometrics*, (2021).

Gordon, Robert James, Kane, Thomas J., and Staiger, Douglas, "Identifying Effective Teachers Using Performance on the Job," (Washington, D.C., The Hamilton Project, 2006).

Government Accountability Office, "Race to the Top: Survey of State Education Agencies' Capacity to Implement Reform," (Washington, D.C., 2015a).

——, "Race to the Top: Survey of School Districts' Capacity to Implement Reform (GAO-15-317SP, April 2015), an E-supplement to GAO-15-295," (2015b).

Gupta, Snigdha, Supplee, Lauren H., Suskind, Dana, and List, John A., "Failed to Scale: Embracing the Challenge of Scaling in Early Childhood," in *The Scale-Up Effect in Early Childhood and Public Policy* (Amsterdam, Routledge, 2021).

Hallgren, Kristin, James-Burdumy, Susanne, and Perez-Johnson, Irma, "State Requirements for Teacher Evaluation Policies Promoted by Race to the Top. NCEE Evaluation Brief. NCEE 2014-4016," (National Center for Education Evaluation and Regional Assistance, 2014).

Herlihy, Corinne, Karger, Ezra, Pollard, Cynthia, Hill, Heather C., Kraft, Matthew A., Williams, Megan, and Howard, Sarah, "State and local efforts to investigate the validity and reliability of scores from teacher evaluation systems," *Teachers College Record*, 116 (2014), 1–28 (SAGE Publications Sage CA: Los Angeles, CA).

Hess, Frederick M., and Bell, Brendan, "Thanks to One Reform, School Principals Spend Weeks Doing Paperwork," *National Review*, <https://www.nationalreview.com/2017/09/education-reform-wrong-way-nevada-principals-spend-weeks-doing-paperwork/> (2017).

Honig, Meredith I., "New directions in education policy implementation: Confronting complexity," (Suny Press, 2006).

Howell, William G., and Magazinnik, Asya, "Presidential Prescriptions for State Policy: Obama's Race to the Top Initiative," *Journal of Policy Analysis and Management*, 36 (2017), 502–531.

——, "Financial Incentives in Vertical Diffusion: The Variable Effects of Obama's Race to the Top Initiative on State Policy Making," *State Politics & Policy Quarterly*, 20 (2020), 185–212 (Cambridge University Press).

Hunter, Seth B., and Springer, Matthew G., "Critical Feedback Characteristics, Teacher Human Capital, and Early-Career Teacher Performance: A Mixed-Methods Analysis," *Educational Evaluation and Policy Analysis*, (2022), 01623737211062913 (American Educational Research Association).

Jackson, C. Kirabo, "What do test scores miss? The importance of teacher effects on non–test score outcomes," *Journal of Political Economy*, 126 (2018), 2072–2107.

Jackson, C. Kirabo, Wigger, Cora, and Xiong, Heyu, "Do school spending cuts matter? Evidence from the great recession," *American Economic Journal: Economic Policy*, 13 (2021), 304–35.

James, Jessalynn, and Wyckoff, James H., "Teacher evaluation and teacher turnover in equilibrium: Evidence from DC public schools," *AERA Open*, 6 (2020), 2332858420932235.

Jepsen, Christopher, and Rivkin, Steven, "Class size reduction and student achievement the potential tradeoff between teacher quality and class size," *Journal of human resources*, 44 (2009), 223–250 (University of Wisconsin Press).

Koedel, Cory, Li, Jiaxi, Springer, Matthew G., and Tan, Li, "Teacher performance ratings and professional improvement," *Journal of Research on Educational Effectiveness*, 12 (2019), 90–115.

Kraft, Matthew A., "Teacher effects on complex cognitive skills and social-emotional competencies," *Journal of Human Resources*, 54 (2019), 1–36.

Kraft, Matthew A., Brunner, Eric J., Dougherty, Shaun M., and Schwegman, David J., "Teacher accountability reforms and the supply and quality of new teachers," *Journal of Public Economics*, 188 (2020), 104212.

Kraft, Matthew A., and Christian, Alvin, "Can teacher evaluation systems produce high-quality feedback? An administrator training field experiment," *American Educational Research Journal*, 59 (2022), 500–537.

Kraft, Matthew A., and Gilmour, Allison, "Revisiting the widget effect: Teacher evaluation reforms and the distribution of teacher effectiveness," *Educational Researcher*, 46 (2017), 234–249.

Lemos, Renata, Muralidharan, Karthik, and Scur, Daniela, "Personnel management and school productivity: Evidence from india," (National Bureau of Economic Research, 2021).

Liebowitz, David D., "Teacher evaluation for accountability and growth: Should policy treat them as complements or substitutes?," *Labour Economics*, (2021), 102024.

List, John A., "The voltage effect: How to make good ideas great and great ideas scale," (Currency, 2022).

Loeb, Susanna, Miller, Luke C., and Wyckoff, James, "Performance screens for school improvement: The case of teacher tenure reform in New York City," *Educational Researcher*, 44 (2015), 199–212.

Manna, Paul, "Collision course: Federal education policy meets state and local realities," (Washington, D.C., CQ Press, 2010).

McGuinn, Patrick, "Stimulating reform: Race to the Top, competitive grants and the Obama education agenda," *Educational Policy*, 26 (2012), 136–159.

——, "From No Child Left Behind to the Every Student Succeeds Act: Federalism and the Education Legacy of the Obama Administration," *Publius: The Journal of Federalism*, 46 (2016), 392–415.

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

Mintrop, Heinrich, and Trujillo, Tina, "The practical relevance of accountability systems for school improvement: A descriptive analysis of California schools," *Educational Evaluation and Policy Analysis*, 29 (2007), 319–352.

Muralidharan, Karthik, and Niehaus, Paul, "Experimentation at scale," *Journal of Economic Perspectives*, 31 (2017), 103–24.

Muralidharan, Karthik, and Singh, Abhijeet, "Improving public sector management at scale? experimental evidence on school governance india," (National Bureau of Economic Research, 2020).

Murnane, Richard J., "US high school graduation rates: Patterns and explanations," *Journal of Economic Literature*, 51 (2013), 370–422.

Murphy, Joseph, Hallinger, Philip, and Heck, Ronald H., "Leading via teacher evaluation: The case of the missing clothes?," *Educational Researcher*, 42 (2013), 349–354.

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

——, "State-by-state evaluation timeline briefs," (2016).

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

——, "Teacher Contract Database," https://www.nctq.org/contract-database/home (2022).

Neal, Derek, and Schanzenbach, Diane Whitmore, "Left behind by design: Proficiency counts and test-based accountability," *The Review of Economics and Statistics*, 92 (2010), 263–283.

Neumerski, Christine M., Grissom, Jason A., Goldring, Ellen, Rubin, Mollie, Cannata, Marisa, Schuermann, Patrick, and Drake, Timothy A., "Restructuring instructional leadership: How multiple-measure teacher evaluation systems are redefining the role of the school principal," *The Elementary School Journal*, 119 (2018), 270–297.

Oyer, P., and Schaefer, S., "Personnel economics: Hiring and incentives," in *Handbook of Labor Economics*, David Card and Orley Ashenfelter, eds. (Amsterdam, Elsevier, 2011).

Paige, Mark, "Moving forward while looking back: Lessons learned from teacher evaluation litigation concerning Value-Added Models (VAMs)," *Education Policy Analysis Archives*, 28 (2020), 64–64.

Papay, John P., Taylor, Eric S., Tyler, John H., and Laski, Mary E., "Learning job skills from colleagues at work: Evidence from a field experiment using teacher performance data," *American Economic Journal: Economic Policy*, 12 (2020), 359–88.

Petek, Nathan, and Pope, Nolan, "The multidimensional impact of teachers on students," *Journal of Political Economy*, Forthcoming (2022).

Phipps, Aaron R., and Wiseman, Emily A., "Enacting the rubric: Teacher improvements in windows of high-stakes observation," *Education Finance and Policy*, 16 (2021), 283–312.

Prendergast, Canice, "The provision of incentives in firms," *Journal of Economic Literature*, 37 (1999), 7–63.

Putnam, Hannah, Ross, Elizabeth, and Walsh, Kate, "Making a Difference: Six Places Where Teacher Evaluation Systems Are Getting Results.," (Washington, D.C., National Council on Teacher Quality, 2018).

Reardon, Sean, Ho, Andrew, Shear, Benjamin, Fahle, Erin, Kalogrides, Demetra, and Chavez, Belen, "Stanford Education Data Archive (Version 4.0)," (2021).

Rodriguez, Luis A., Swain, Walker A., and Springer, Matthew G., "Sorting through performance evaluations: The influence of performance evaluation reform on teacher attrition and mobility," *American Educational Research Journal*, 57 (2020), 2339–2377.

Rothstein, Jesse, "Teacher quality policy when supply matters," *American Economic Review*, 105 (2015), 100–130.

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

Sartain, Lauren, and Steinberg, Matthew P., "Teachers' Labor Market Responses to Performance Evaluation Reform: Experimental Evidence from Chicago Public Schools," *Journal of Human Resources*, 51 (2016), 615–655.

——, "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, 2021).

Sawchuk, Stephen, "Teacher Evaluation Heads to the Courts," *Education Week* (2015) (Oct. 5, 2021).

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

Springer, Matthew G., "The influence of an NCLB accountability plan on the distribution of student test score gains," *Economics of Education Review*, 27 (2008), 556–563 (Elsevier).

——, "Performance Incentives: Their Growing Impact on American K-12 Education," (Brookings Institution Press, 2010).

Staiger, Douglas O., and Rockoff, Jonah E., "Searching for effective teachers with imperfect information," *Journal of Economic Perspectives*, 24 (2010), 97–118.

Stecher, Brian M., Holtzman, Deborah J., Garet, Michael S., Hamilton, Laura S., Engberg, John, Steiner, Elizabeth D., Robyn, Abby, Baird, Matthew D., Gutierrez, Italo A., Peet, Evan D., Brodziak de los Reyes, Iliana, Fronberg, Kaitlin, Weinberger, Gabriel, Hunter, Gerald P., and Chambers, Jay, "Improving Teaching Effectiveness: Final Report: The Intensive Partnerships for Effective Teaching Through 2015–2016," (Washington, D.C., RAND, 2018).

Steinberg, Matthew P., and Sartain, Lauren, "Does teacher evaluation improve school performance? Experimental evidence from Chicago's Excellence in Teaching project," *Education Finance and Policy*, 10 (2015), 535–572.

Sun, Liyang, and Abraham, Sarah, "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects," *Journal of Econometrics*, 225 (2021), 175–199 (Elsevier).

Taylor, Eric S., "Employee evaluation and skill investments: Evidence from public school teachers," *EdWorkingPapers.com* (Annenberg Institute at Brown University, 2022).

Taylor, Eric S., and Tyler, John H., "The effect of evaluation on teacher performance," *The American Economic Review*, 102 (2012), 3628–3651.

U.S. Bureau of Economic Analysis, "CAGDP2 Gross domestic product (GDP) by county and metropolitan area," (2021).

U.S. Department of Education, "Common Core of Data," (2021).

——, "Title II," <https://title2.ed.gov/Public/Home.aspx> (2022) (Dec. 9, 2022).

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

Wong, Kenneth K., "Federal ESEA waivers as reform leverage: Politics and variation in state implementation," *Publius: The Journal of Federalism*, 45 (2015), 405–426 (Oxford University Press).

Wooldridge, Jeffrey M., "Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators," *Available at SSRN 3906345*, (2021).

Woulfin, Sarah L., and Rigby, Jessica G., "Coaching for coherence: How instructional coaches lead change in the evaluation era," *Educational Researcher*, 46 (2017), 323–328.

Zhou, Jin, Baulos, Alison, Heckman, James J., and Liu, Bei, "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*, (2021).

TABLE I

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	61.6	6.62	572	ACS
College Enrollment	62.5	6.03	572	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 II

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	0.5991	0.0071	0.2503	0.2462
	(0.8062)	(0.7359)	(0.6044)	(0.6026)
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	571	571	571	571

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, NAEP Math, NAEP ELA, and either baseline High School Graduation or Baseline College Enrollment. Standard errors are clustered by state. *p < 0.05, ** p <0.01, ***p <0.001.

TABLE III

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	2.0040	1.9792*	-0.4238	-0.8849
	(1.1312)	(0.9374)	(0.9212)	(0.8977)
Teacher Evaluation X Index	-0.2524	-0.3456*	0.1211	0.1971
	(0.1574)	(0.1493)	(0.1532)	(0.1565)
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	571	571	571	571

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 design features we observe in our index. *p < 0.05, ** p <0.01, ***p <0.001.

TABLE IV

Effect of Teacher Evaluation on Working Conditions

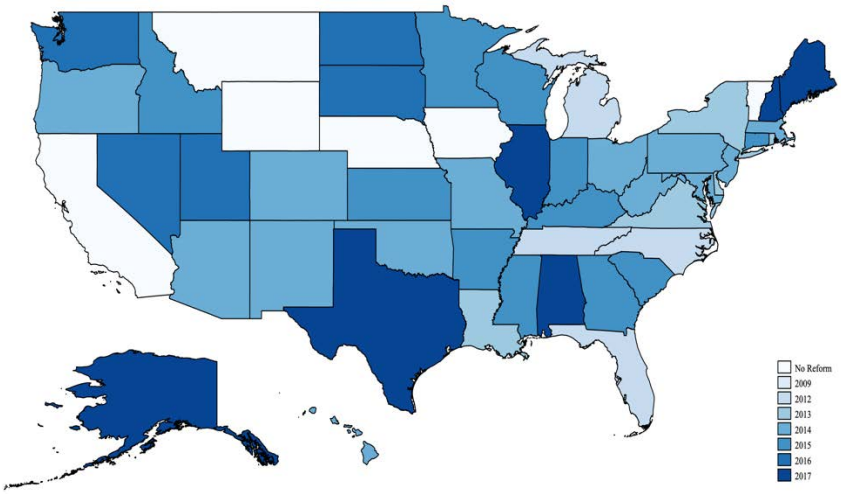
	(1)	(2)	(3)
Outcome	Pay Satisfaction	Job Security	Job Satisfaction
Tch Eval	-0.057**	-0.071**	-0.029*
	(0.015)	(0.027)	(0.014)
Mean 2003	0.073	0.071	0.031
n	150,930	150,930	150,930
Outcome	Control over Materials	Control over Content	Control over Technique
Tch Eval	-0.021	-0.009	0.004
	(0.016)	(0.0250)	(0.004)
Mean 2003	0.320	0.362	0.706
n	150,930	150,930	150,930

Note: Pay Satisfaction is a binary outcome equal to one if a teacher strongly disagreed, "If I could get a higher paying job I'd leave teaching as soon as possible" and zero otherwise. Job Security is a binary outcome equal to one if a teacher strongly disagreed, "I worry about the security of my job because of the performance of my students or my school on state and/or Local tests." Job Satisfaction is a binary outcome equal to one if a teacher strongly agreed that they were, "generally satisfied with being a teacher at this school" and zero otherwise. Control over Materials is a binary variable equal to one if teachers indicated they had a great deal of control over "selecting content, topics, and skills to be taught "when asked, "How much actual control do you have IN YOUR CLASSROOM at this school over the following areas of your planning and teaching?" Control over Content is a binary variable equal to one if teachers indicated they had a great deal of control over "selecting textbooks and other instructional materials" when asked, "How much actual control do you have IN YOUR CLASSROOM at this school over the following areas of your planning and teaching?" Control over Content is a binary variable equal to one if teachers indicated they had a great deal of control over "selecting textbooks and other instructional materials" when asked, "How much actual control do you have IN YOUR CLASSROOM at this school over the following areas of your planning and teaching?" Control over Technique is a binary variable equal to one if teachers indicated they had a great deal of control over "selecting teaching techniques" when asked, "How much actual control do you have IN YOUR CLASSROOM at this school over the following areas of your planning and teaching?" Control over Technique, Control over Content, Control over Technique are binary outcomes equal to one if teacher reported they had a "great deal of actual control" over either inst, when asked, "How much actual control do you have IN YOUR CLASSROOM at this school over the following areas of your planning and teaching?"Models include data from 4 years: The public teacher Schools and Staffing survey in 2003, 2007, and 2011; the public teacher National Teacher Principal Survey in 2015. All models include state and year fixed effects. Standard errors are clustered by state. Samples sizes rounded in accordance with NCES restricted use rules. Models estimated using teacher-level inverse probability weights. Standard errors are clustered by state. *p < 0.05, ** p <0.01, ***p <0.001.

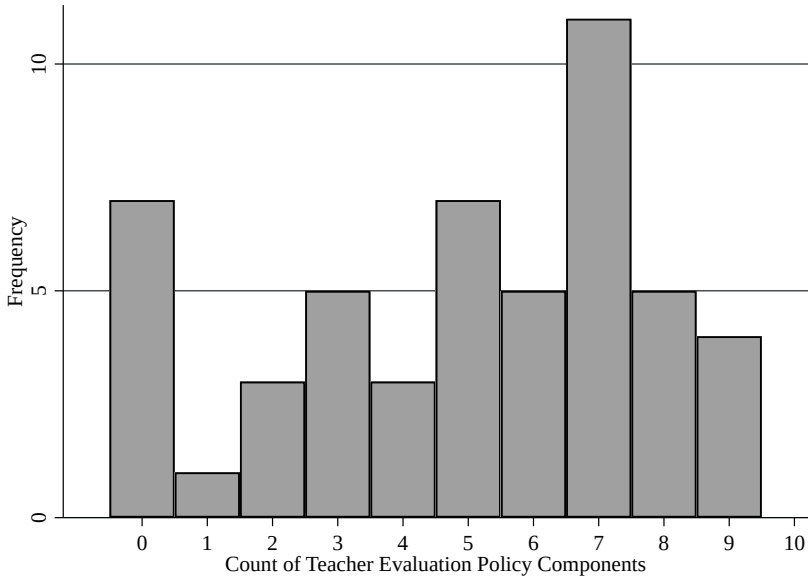
FIGURE I

Teacher Evaluation Implementation

Panel A. State Implementation Map



Panel B. Histogram of Teacher Evaluation Reform Quality Index

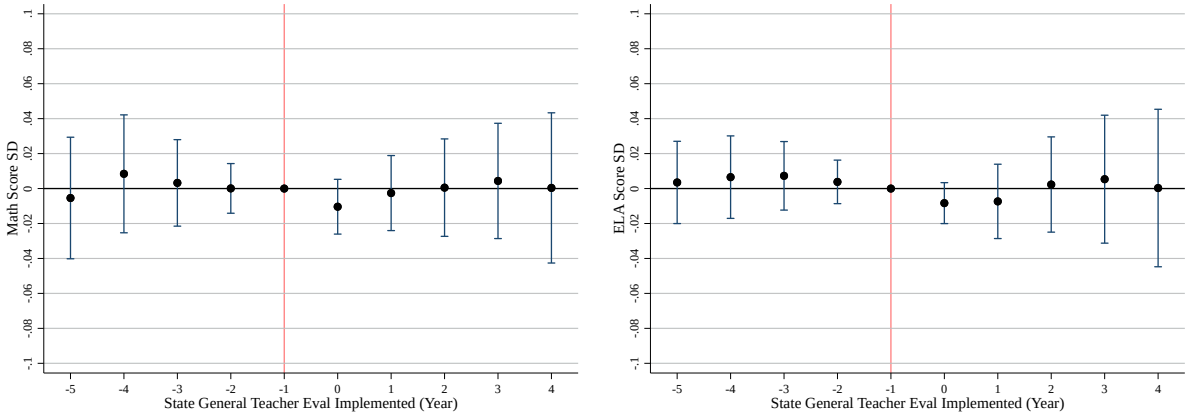


Note: The index for comparison states is zero even if they implemented a design feature of teacher evaluation reform. See Appendix B for details on the design features of the index. All years are the spring of the school year.

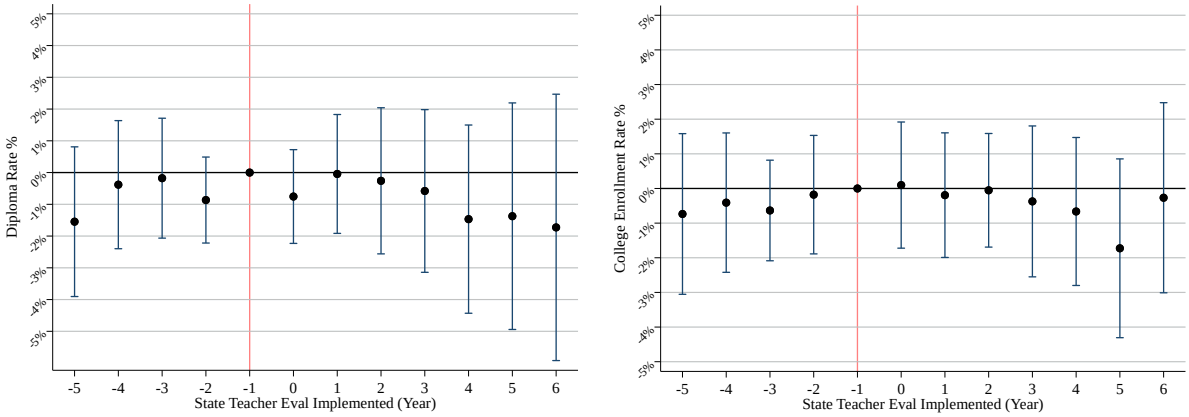
FIGURE II

Event Study: Effects on Achievement and Attainment

Panel A. Achievement



Panel B. Attainment

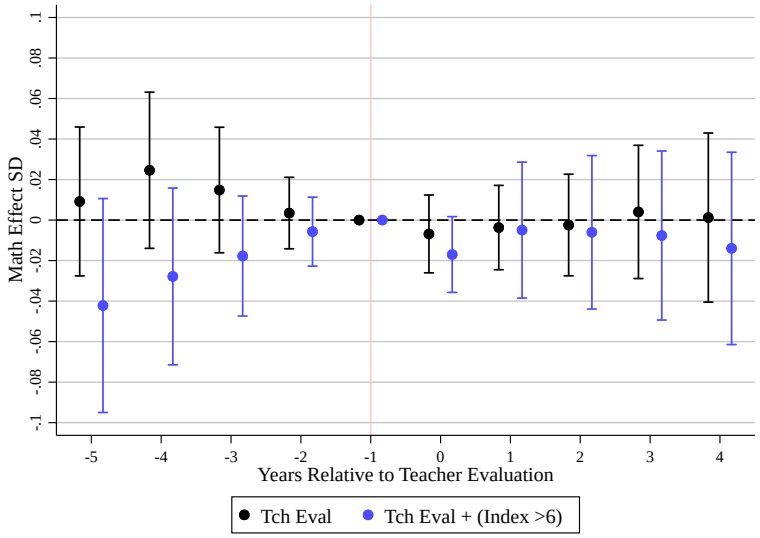


Note: Models with achievement outcomes include district fixed effects, grade fixed effects, year fixed effects, and baseline covariates measured in 2009 interacted with a linear year trend: percent Black, percent Hispanic, percent Native American, percent Asian, total enrollment, urban/city, GDP, poverty 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.

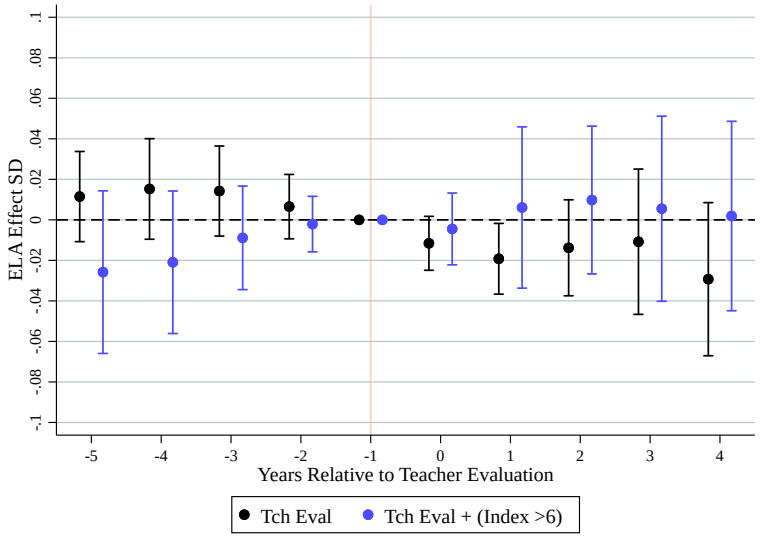
FIGURE III

Event Study: Heterogeneity by Index

Panel A. Math



Panel B. ELA

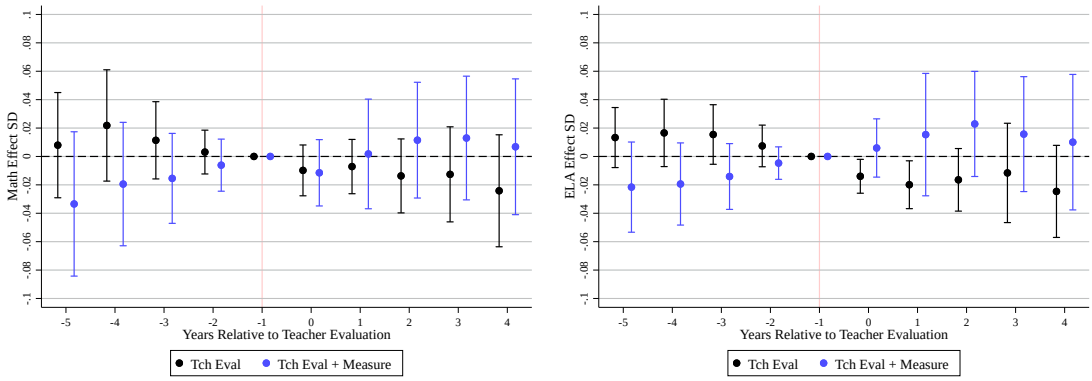


Note: Models include two estimates for each relative time period: the main event study dummies and a set of event study dummies interacted with a time-invariant indicator equal to one for states that had an index from 7 to 10. The black estimates "Tch Eval" are the main event study dummies. The blue 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.

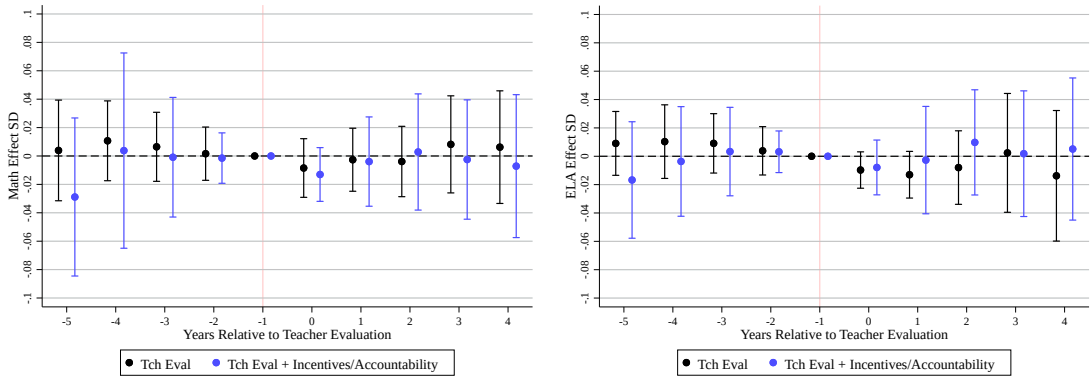
FIGURE IV

Event Study: Heterogeneity by System Design

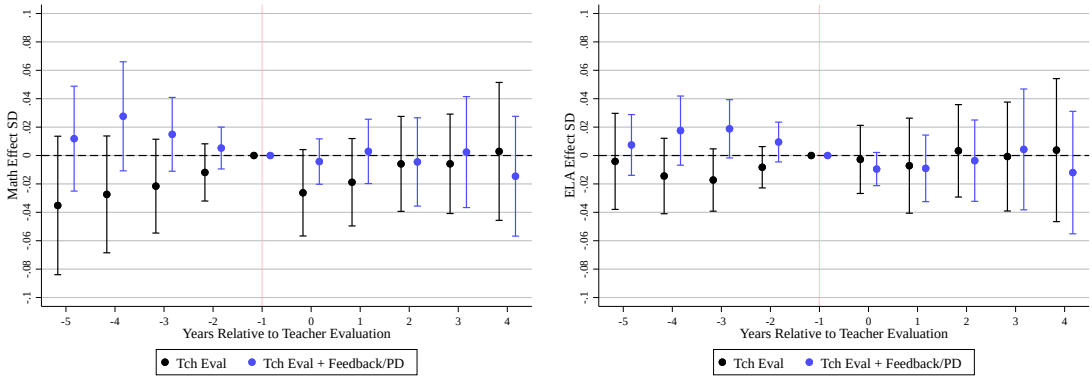
Panel A. Measurement



Panel B. Accountability and Incentives



Panel B. Feedback and Professional Development

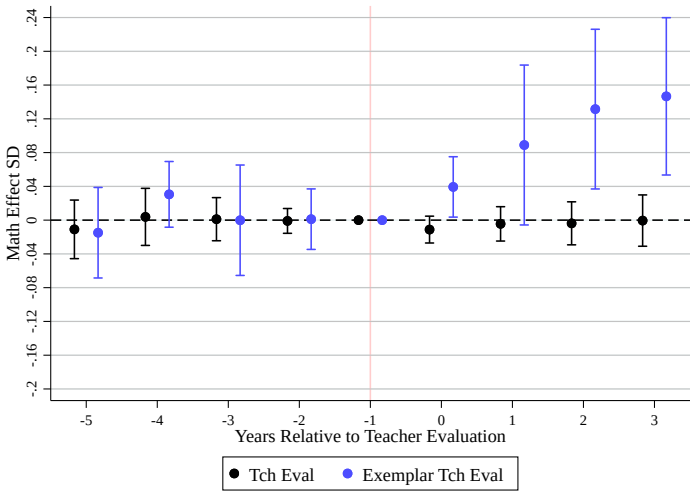


Note: Models include the main event study dummies and a set of event study dummies interacted with a time-invariant indicator equal to one for a specified system design. The black estimates are the main event study dummies. The blue estimates are the linear combination of the estimates for the "high 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.

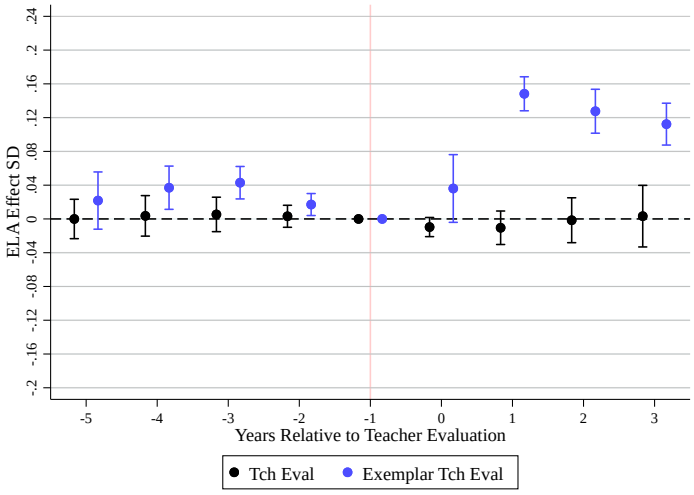
FIGURE V

Event Study: Heterogeneity by Exemplar Evaluation Systems

Panel A. Math



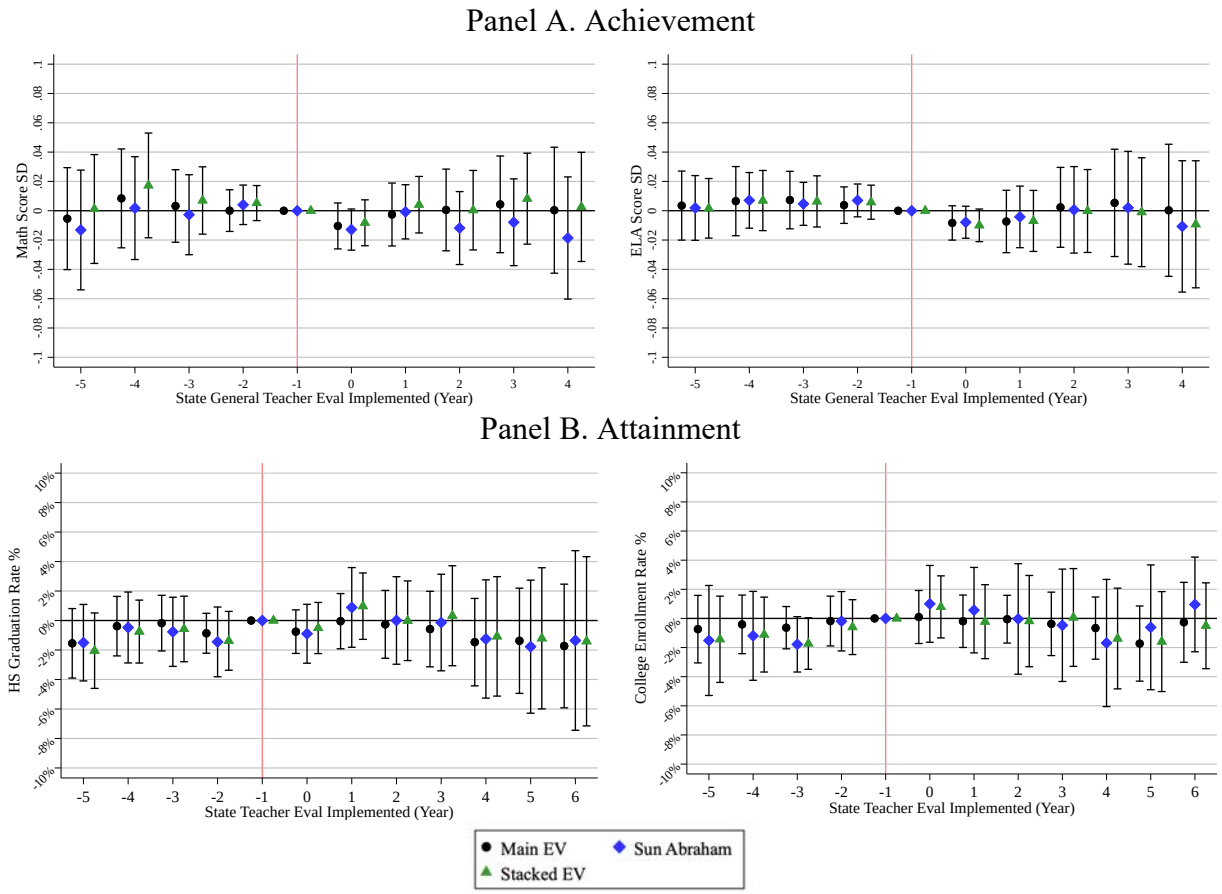
Panel B. ELA



Note: Exemplar teacher evaluation systems include: DISD, DCPS, DPS, NPS, TN, NM. 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 VI

Event Study and Estimates 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.

APPENDIX TABLE A.1

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	
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 A.2

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. *p < 0.05, ** p <0.01, ***p <0.001.

APPENDIX TABLE A.3

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	-0.7361	-1.5471
	(1.1545)	(1.1736)
-4 Pre	-0.4086	-0.3813
	(1.0014)	(1.0044)
-3 Pre	-0.6350	-0.1772
	(0.7227)	(0.9400)
-2 Pre	-0.1784	-0.8656
	(0.8513)	(0.6742)
0 Post	0.0978	-0.7548
	(0.9065)	(0.7349)
1 Post	-0.1929	-0.0437
	(0.8953)	(0.9320)
2 Post	-0.0528	-0.2599
	(0.8163)	(1.1457)
3 Post	-0.3729	-0.5801
	(1.0843)	(1.2754)
4 Post	-0.6640	-1.4668
	(1.0633)	(1.4758)
5 Post	-1.7264	-1.3742
	(1.2846)	(1.7762)
6 Post	-0.2672	-1.7268
	(1.3660)	(2.0883)
n	571	571

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. *p < 0.05, ** p <0.01, ***p <0.001.

APPENDIX TABLE A.4

Effect of Teacher Evaluation With Alternative Covariates

	(1)	(2)	(3)	(4)
	Math	Math	ELA	ELA
Teacher Evaluation	-0.0173	-0.0193	-0.0180	-0.0162
	(0.0126)	(0.0129)	(0.0103)	(0.0099)
District FE	X	X	X	X
Grade FE	X	X	X	X
Year FE	X	X	X	X
District Time-Varying Controls	X	X	X	X
District Baseline Ach Controls		X		X
n	460,401	460,401	491,944	491,944
Outcome	HS Grad	HS Grad	College Enroll	College Enroll
Teacher Evaluation	0.6084	0.2601	0.3081	0.5672
	(0.7482)	(0.7511)	(0.5980)	(0.5845)
State FE	X	X	X	X
Year FE	X	X	X	X
State Time-Varying Controls	X	X	X	X
State Baseline Attain Controls		X		X
n	571	571	571	571
	Math	Math	ELA	ELA
Teacher Evaluation	-0.0196	-0.0149	-0.0176	-0.0160
	(0.0130)	(0.0126)	(0.0105)	(0.0108)
District FE	X	X	X	X
Grade FE	X	X	X	X
Year FE	X	X	X	X
State Time-Varying Controls	X	X	X	X
State Baseline Ach Controls		X		X
n	460,401	460,401	491,944	491,944

Note: District-time varying controls include Percent Black, Percent Hispanic, Percent Native American, Percent Asian, Total Enrollment, Urban/City, GDP, Poverty Index, Unemployment Rate. District Baseline Ach Controls include ELA Score, and Math Score interacted with a linear time trend. State Time-Varying controls include Percent Black, Percent Hispanic, Percent Native American, Percent Asian, Total Enrollment, Urban/City, GDP, Poverty Index, Unemployment Rate. State Baseline Attain controls include baseline high school graduation and college enrollment interacted with a linear time trend. State Baseline Ach Controls include NAEP Math and NAEP ELA. We exclude Student-Teacher Ratio and Per Pupil expenditures—which we include in the main specification—because both controls are likely endogenous. Standard errors are clustered by state. *p < 0.05, ** p <0.01, ***p <0.001.

APPENDIX TABLE A.5

Effect of Teacher Evaluation on College Graduation

	(1)	(2)
Teacher Evaluation	0.5991	0.0071
	(0.8062)	(0.7359)
State FE	X	X
Year FE	X	X
State Ed Controls		X
State SES Controls		X
Attainment Controls		X
n	571	571

Note: College graduation is calculated using the percent of 24-year-olds who ever enrolled in a college in a given year by state of birth, again using PUMS person weights from 2008 to 2020. See notes in Table 2 for a full list of covariates. Standard errors are clustered by state. *p < 0.05, ** p <0.01, ***p <0.001.

APPENDIX TABLE A.6

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	0.7269	0.2987	0.3058	0.2665
	(0.8750)	(0.7436)	(0.5991)	(0.5831)
Eval X High Quality	-0.2923	-0.6899	-0.1269	-0.0469
	(0.7897)	(0.7566)	(0.7168)	(0.7062)
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	571	571	571	571

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. *p < 0.05, ** p <0.01, ***p <0.001.

APPENDIX TABLE A.7

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	0.4184	0.5191	1.0059	0.3383	-0.0790	-0.8813
	(0.6662)	(0.8095)	(0.8826)	(0.5741)	(0.6277)	(0.7022)
Eval X Measurement	-1.1096			-0.2381		
	(0.9357)			(0.8269)		
Eval X Incent/Account		-1.2575			0.7735	
		(0.7384)			(0.7291)	
Eval X Feedback/PD			-1.4558			1.6165*
			(0.8036)			(0.6041)
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	571	571	571	571	571	571

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. *p < 0.05, ** p <0.01, ***p <0.001.

APPENDIX TABLE A.8

Effects of Exemplar Evaluation Systems

	(1)	(2)	(3)	(4)
Outcome	Math	Math	ELA	ELA
Teacher Evaluation	-0.0220	-0.0105	-0.0247*	-0.0120
	(0.0125)	(0.0114)	(0.0118)	(0.0094)
Exemplar	0.0855	0.0925	0.0544*	0.0702*
	(0.0549)	(0.0527)	(0.0256)	(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 teacher evaluation systems include: DISD, DCPS, DPS, NPS, TN, NM. 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. *p < 0.05, ** p <0.01, ***p <0.001.

APPENDIX TABLE A.9

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	-0.1111	-0.0551	-0.0230
	(0.7015)	(0.7121)	(0.7214)
Teacher Evaluation X Percent FRPL	-1.5870*		
	(0.6810)		
Teacher Evaluation X Percent Black		-0.6341	
		(0.8794)	
Teacher Evaluation X Percent Hispanic			-0.9747
			(0.6308)
n	571	571	571
Panel D. College Enrollment			
Teacher Eval	0.2707	0.3101	0.2512
	(0.6033)	(0.6112)	(0.6017)
Teacher Evaluation X Percent FRPL	0.4178		
	(0.7247)		
Teacher Evaluation X Percent Black		0.7170	
		(0.6072)	
Teacher Evaluation X Percent Hispanic			0.4517
			(0.5436)
n	571	571	571

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. *p < 0.05, ** p <0.01, ***p <0.001.

APPENDIX TABLE A.10

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	-0.2778	0.0685
	(0.8994)	(0.9543)
State FE	X	X
Year FE	X	X
State-Specific Trends	X	X
n	571	571

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. *p < 0.05, ** p <0.01, ***p <0.001.

APPENDIX TABLE A.11

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	0.5176	0.1344	-0.6745	-0.6610
	(0.6653)	(0.6490)	(0.6216)	(0.6131)
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	522	522	522	522

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). *p < 0.05, ** p <0.01, ***p <0.001.

APPENDIX TABLE A.12

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. *p < 0.05, ** p <0.01, ***p <0.001.

APPENDIX TABLE A.13

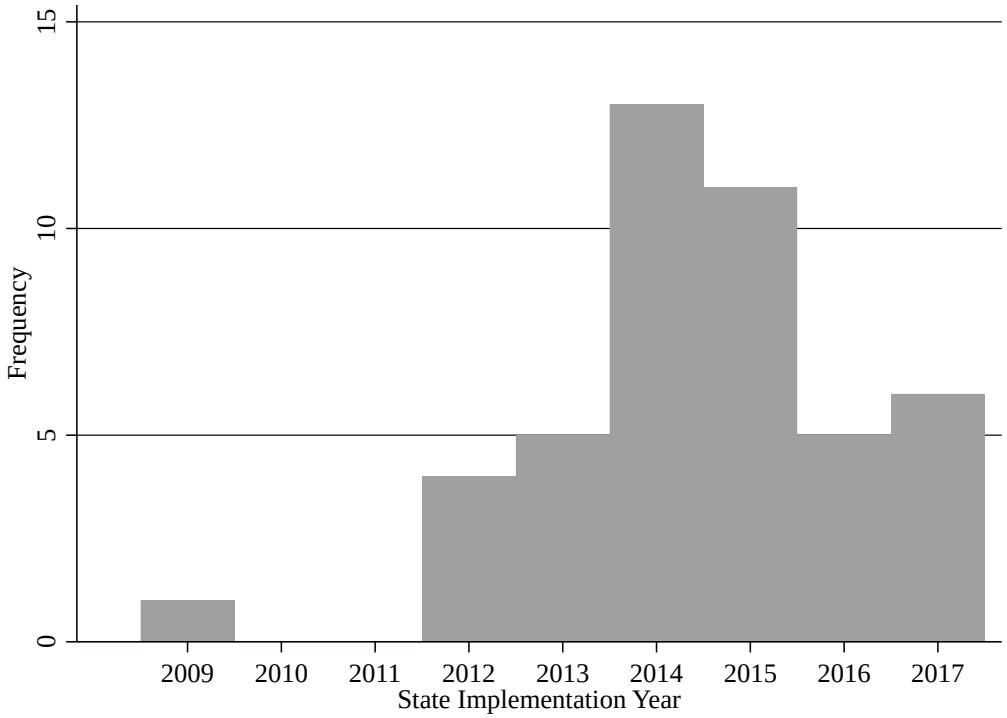
Teacher Evaluation Effect on Salary

	(1)	(2)
Outcome	Salary	Salary
Tch Eval	-871.2168	-251.1776
	(501.0239)	(366.1084)
State FE	X	X
Year FE	X	X
State Ed Controls		X
State SES Controls		X
n	571	571

Note: Salary is real average teacher salaries in 2020 dollars caculated using the ACS. Teacher supply is the proportion of adults (age 18 to 65) in a state who earned either traditional or alternative teacher certifications (U.S. Department of Education 2022).

APPENDIX FIGURE A.1

Implementation of Evaluation Reforms by Year

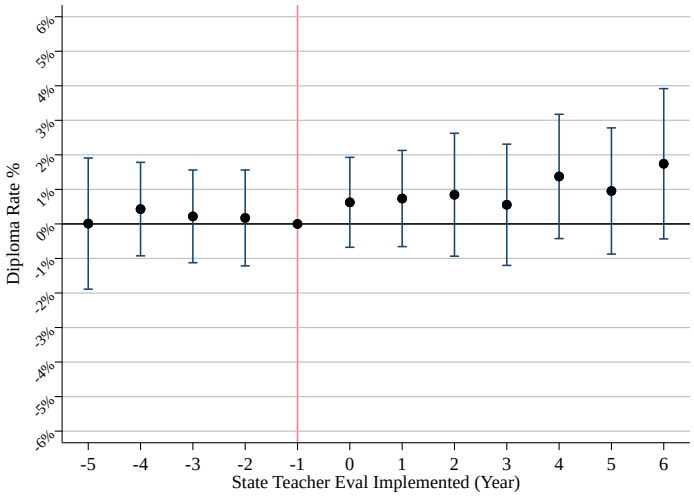


Note: All years are the spring of the school year.

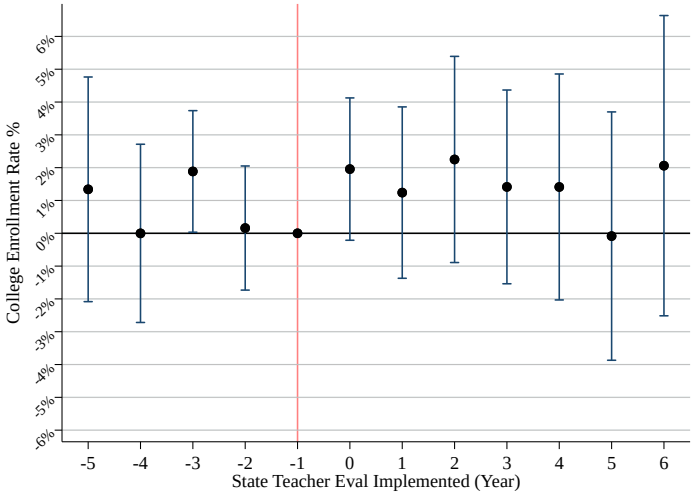
APPENDIX FIGURE A.2

Event Study: Attainment Measured at Age 20

Panel A. Diploma Rate



Panel B. College Enrollment Rate

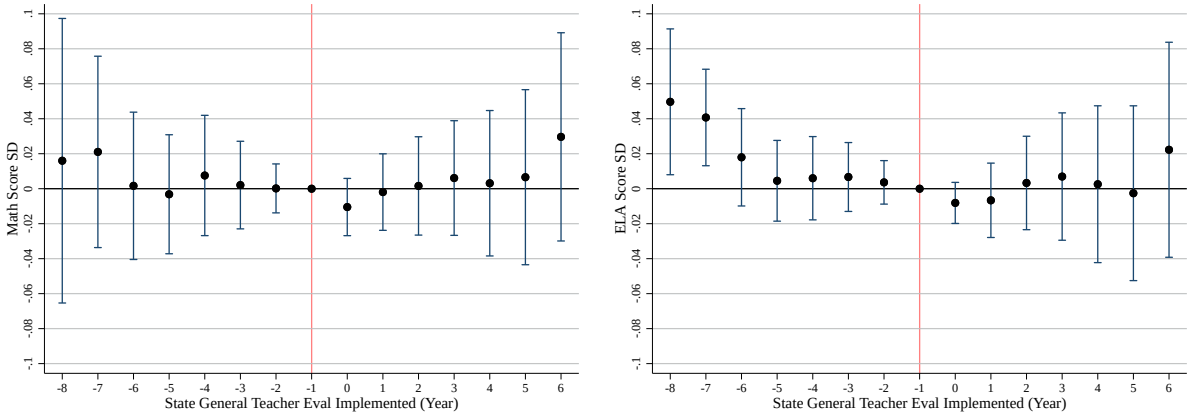


Note: Models include state fixed effects, year fixed effects, and baseline covariates measured in 2009 interacted with a linear year trend: percent Black, percent Hispanic, percent Native American, percent Asian, total enrollment, urban/city, GDP, percent FRPL, unemployment rate, student teacher ratio,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

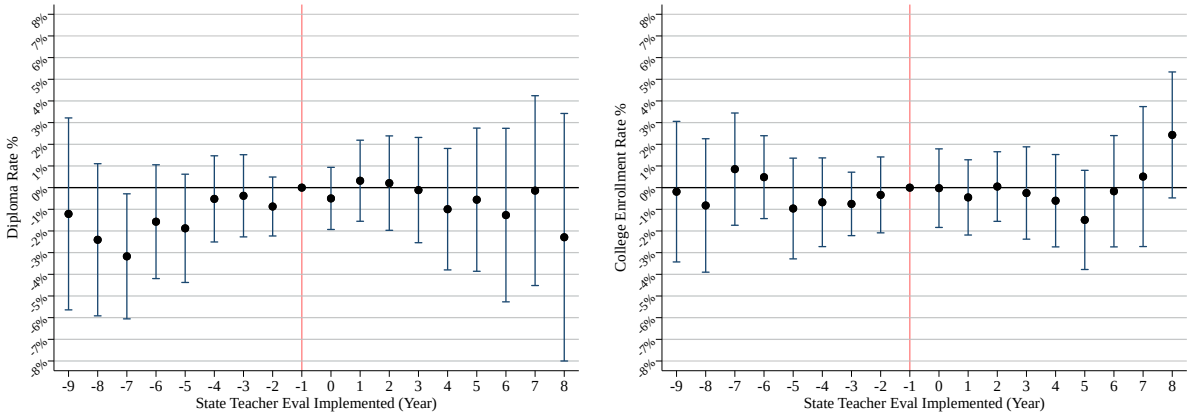
APPENDIX FIGURE A.3

Event Study: Untrimmed Effects on Achievement and Attainment

Panel A. Achievement



Panel B. Attainment

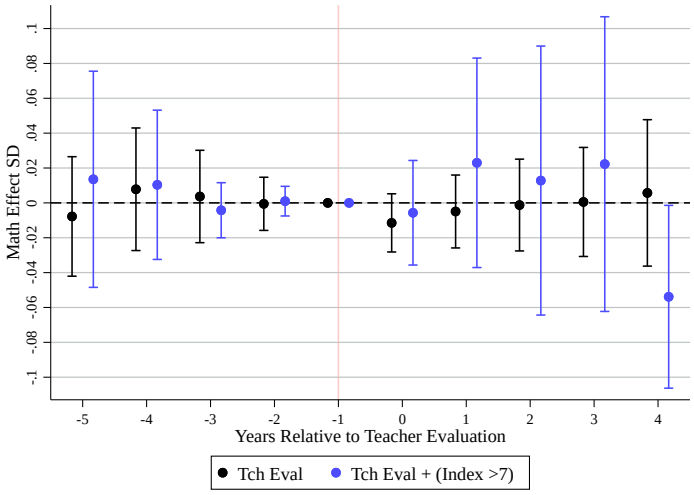


Note: See notes in Table 2 for a full list of covariates. Standard errors are clustered by state. Event studies describe the untrimmed sample.

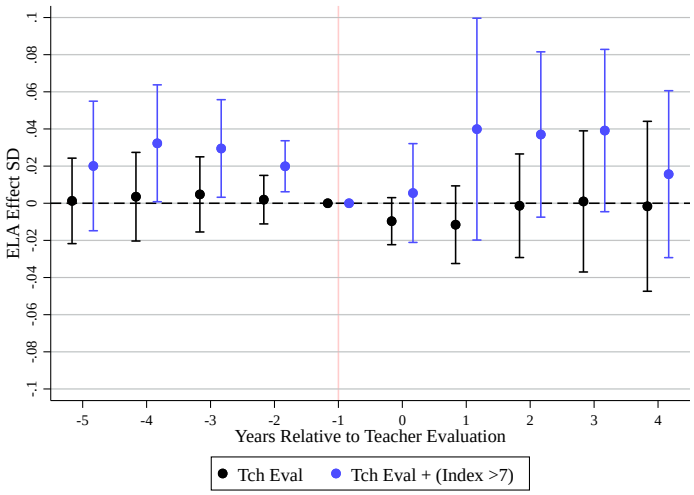
APPENDIX FIGURE A.4

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 B.1

Teacher Evaluation Reform Design Features

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 feature of teacher evaluations.	Doherty & Jacobs (2015)	27
Measurement	Student Survey	Student surveys are a required feature 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 feature of teacher evaluations.	Bleiberg & Harbatkin (202); 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 (NCTQ 2011, 2019; Doherty and Jacobs 2015).

APPENDIX TABLE B.2

State Teacher Evaluation Feature 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

APPENDIX TABLE B.3

Teacher Evaluation Categorical Constructs and Quality Measures

Category	**Descriptions**	**State #**
Measurement	Teacher evaluation systems include at least three of the following features: (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 features: (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