Journal of Public Economics 188 (2020) 104212

Contents lists available at ScienceDirect

Journal of Public Economics

journal homepage: www.elsevier.com/locate/jpube

Teacher accountability reforms and the supply and quality of new teachers☆

Matthew A. Kraft [a,*], Eric J. Brunner [b], Shaun M. Dougherty [c], David J. Schwegman [d]

[a] Brown University, United States of America
[b] University of Connecticut, United States of America
[c] Vanderbilt University, United States of America
[d] American University, United States of America

ARTICLE INFO

Article history:
Received 14 February 2019
Received in revised form 16 May 2020
Accepted 18 May 2020
Available online 23 July 2020

JEL classifications:
I21
I28
J22
J24

Keywords:
Teacher supply
Teacher evaluation
Teacher quality
Education
Teacher hiring

ABSTRACT

In recent years, states have sought to increase accountability for public school teachers by implementing a package of reforms centered on high-stakes evaluation systems. We examine the effect of these reforms on the supply and quality of new teachers. Leveraging variation across states and time, we find that accountability reforms reduced the number of newly licensed teacher candidates and increased the likelihood of unfilled teaching positions, particularly in hard-to-staff schools. Evidence also suggests that reforms increased the quality of newly hired teachers by reducing the likelihood new teachers attended unselective undergraduate institutions. Decreases in job security, satisfaction, and autonomy are likely mechanisms for these effects.

1. Introduction

In response to mounting public pressure and strong incentives from the federal government, state legislatures across the country have enacted laws aimed at increasing accountability for public school teachers. By 2016, 44 states had implemented major reforms to their teacher evaluation systems intended to increase the rigor of performance reviews and streamline the dismissal process. A handful of states also enacted laws that restricted tenure protections, lengthened the probationary period for teachers, limited the scope of collective bargaining with teacher unions, and eliminated mandatory union dues. Together, these accountability reforms weakened the employment protections unions could provide to teachers.

Proponents of accountability reforms argued that high-stakes evaluation systems combined with merit pay could transform the teacher labor force by removing low-performing teachers and attracting higher quality candidates into the profession (Hanushek, 2009; Klein et al., 2010). Opponents argued that high-stakes teacher evaluation systems were unreliable and would only serve to make teaching a less attractive profession (Fullan, 2011). However, limited empirical evidence exists on whether and how these reforms actually have affected the teacher labor market.

What evidence we have comes from studies that examine the effect of accountability reforms on the effort and career decisions of *current* teachers. Prior research has shown that school-level accountability reforms decrease teachers' perceptions of job security (Reback et al., 2014) and increase their effort via reduced absences (Jacob, 2013; Gershenson, 2016). Several studies have also shown that high-stakes evaluation systems have increased voluntary attrition among lower-performing teachers in large urban school districts (Dee and Wyckoff, 2015; Loeb et al., 2015; Sartain and Steinberg, 2016; Cullen et al., 2019).

This paper addresses the largely unexamined question of how accountability reforms affect *new* teachers. Using both event study and difference-in-differences methods, we exploit arguably exogenous variation in the timing of teacher evaluation and other accountability reforms across states to provide the first empirical evidence on how these reforms affected the supply of prospective public school teachers and the ability of schools to fill vacant teaching positions. We then

☆ This work was generously supported by the Spencer Foundation [Award #201700052] and the Institute for Education Sciences [Award # R305A170053].

* Corresponding author at: Brown University, PO Box 1938, Providence, RI 02478, United States of America.
E-mail address: mkraft@brown.edu (M.A. Kraft).

https://doi.org/10.1016/j.jpubeco.2020.104212
0047-2727/

examine the distributional effects of accountability reforms on the quality of newly hired teachers as measured by the selectivity of their undergraduate institutions and whether they held an emergency license.

Several studies have examined the potential learning gains from dismissing low-performing teachers through simulation analyses. These studies implicitly assume that dismissed teachers can always be replaced with average-quality novice teachers (Gordon et al., 2006; Hanushek, 2009; Staiger and Rockoff, 2010; Winters and Cowen, 2013a, 2013b; Goldhaber and Hansen, 2010). However, both qualitative case studies (Johnson, 2019) and simulation analyses (Rothstein, 2015) suggest this assumption may not be realistic, particularly for low-performing and hard-to-staff schools. Our analyses provide a direct empirical test of the assumption that the supply of average-quality novice teachers is relatively inelastic.

We examine the effect of accountability reforms on the supply of new teachers, teaching position vacancies, new teacher quality, and a range of potential mechanisms using a combination of state-by-year panel data from 2002 to 2016 and repeated waves of a nationally representative dataset on public schools and teachers. We model the effects of implementing high-stakes teacher evaluation systems independently as well as part of the joint effect of a bundle of accountability reforms to address the challenge of definitively isolating the effect of evaluation from other related accountability reforms.

We find that implementing high-stakes evaluation reforms reduced the supply of newly licensed teachers by 16–18% and decreased the number of degree completers from graduate teacher preparation programs by 8–10%. The more accountability reforms enacted by states, the greater the intensity of these effects. Flexible models suggest that evaluation reforms resulted in a steady decline in new labor supply over time. We also show that the effects of evaluation persist even when we control for other accountability reforms or restrict the sample of treated states to those where the only accountability reform was the adoption of a new teacher evaluation system. Our findings suggest that new evaluation systems were likely the driving factor behind accountability reforms, but that other contemporaneous accountability reforms intensified effects by further limiting job protections provided by unions.

Given that the number of graduates from teacher preparation programs each year has historically been more than double the number of vacant teaching positions in the U.S. (Cowan et al., 2016), this reduction in teacher supply could have little effect on the ability of schools to fill vacant teaching positions. However, we find that reductions in new teacher supply caused by evaluation reforms appear to bind for schools. Evaluation reforms increased the probability that a school had at least one unfilled teaching vacancy by 2.6 percentage points relative to a pre-reform mean of 4%. As prior evidence would suggest, effects are concentrated in traditionally hard-to-staff schools that serve larger proportions of disadvantaged students (Steele et al., 2010; Clotfelter et al., 2008).

At the same time, we also find suggestive evidence that evaluation reforms increased the quality of new teachers, as measured by the selectivity of teachers' undergraduate institutions. Although this input measure of quality is only a weak proxy for teacher performance, it allows us to explore the nature of supply-side responses to accountability reforms. We find that the increase in teacher quality is primarily driven by a decrease in the supply of teachers coming from less competitive undergraduate institutions.

We explore a range of alternative explanations for the effects we find on teacher supply and quality. These robustness tests suggest that our results are not driven by declines in demand for public school teachers, changes in macroeconomic conditions that affected the broader labor market, or shifts in the demographic composition of school-age children. Finally, we test possible mechanisms for our findings by analyzing new teachers' perceptions of their working conditions. These analyses show that among new teachers, evaluation reforms substantially decreased perceptions about job security, job satisfaction, cooperative effort, and control over their teaching. We conclude with a discussion of the implications for policy, practice, and future research.

2. Teacher accountability reforms

2.1. Teacher evaluation

Efforts to introduce greater accountability in schools and classrooms have ebbed and flowed throughout the history of U.S. public education. The No Child Left Behind (NCLB) Act, passed in 2002, expanded test-based school accountability policies nationally and aimed to improve teacher quality by establishing more rigorous teacher licensure standards. In the following years, a growing body of research on teacher effectiveness exploited new administrative datasets linking students to teachers that NCLB helped to create. Three seminal findings from this research served as signposts for the Obama administration's efforts to promote teacher accountability reform: 1) the effects teachers have on student learning are large and vary considerably across teachers, 2) teachers' paper qualifications are only weakly related to student learning, and 3) teacher evaluation systems were failing to differentiate among teachers despite the large differences in teacher effectiveness (Kraft, 2018).

Starting in 2009, the Obama administration leveraged $4.35 billion from the American Reinvestment and Recovery Act to fund the Race to the Top (RTT) grant competition. The RTT application rubric detailed specific evaluation system reforms required for a competitive proposal. These reforms included evaluating teachers using multiple measures such as student achievement growth, rating teachers on a scale with multiple categories, conducting annual evaluations, and using evaluation ratings to inform high-stakes personnel decisions. In 2012, the Obama administration made adopting evaluation reforms one of several conditions for states to receive a waiver from the increasingly stringent accountability consequences of failing to meet Annual Yearly Progress targets set by NCLB. Between February 2012 and April 2014, forty-three states and D.C. were granted a waiver from NCLB's provisions.

By 2016, a total of 44 states had passed legislation that mandated major teacher evaluation reforms (National Counsel on Teacher Quality [NCTQ], 2016). The design of new evaluation systems differed meaningfully across and within states, but all shared several common features. The vast majority of states adopted state-wide systems but allowed districts discretion to adapt systems to local contexts, while a handful mandated the use of state-designed systems or only provided broad guidelines (Steinberg and Donaldson, 2016). The two most frequently used performance measures were classroom observation ratings and test-based scores, such as value-added measures or student growth percentiles. At the same time, the weights assigned to test-based measures of student performance ranged considerably across states from 0 to 50%. Approximately half of the states phased in the use of test-score based measures across several years following state-wide implementation of the evaluation system.

On paper, most states and districts emphasized that the primary goal of evaluation reforms was to support teachers to improve their instruction. Consistent with this espoused goal, Steinberg and Donaldson (2016) found that the vast majority (83%) of states explicitly linked teacher evaluation ratings to professional development requirements. About 60% of states established systems where low ratings could lead to teachers being dismissed, and almost 50% of states permitted teachers to be denied tenure based on their performance. Some states also delayed the use of evaluation scores to inform high-stakes personnel decisions.

On December 10th of 2015, Present Obama signed the Every Student Succeeds Act (ESSA) into law, which greatly curtailed the degree to which the federal government could mandate or make federal funding conditional on a range of education policy prescriptions, including teacher evaluation reforms. Relevant to our study, many states took advantage of this increased autonomy to revise their teacher evaluation

systems. States first submitted proposed ESSA plans to the Federal Department of Education (ED) in the spring and fall of 2017. Thus, our panel, which runs through 2016, largely captures the effects of high-stakes teacher evaluation reforms implemented before states made their proposed revisions public. By 2019, most states had retained the core features of high-stakes teacher evaluation systems, while 11 states had replaced objective measures of student growth with teacher-developed Student Learning Objectives (SLOs) or abandoned test-based measures altogether (Ross and Walsh, 2019).

2.2. Contemporaneous accountability reforms

Several states also adopted laws that weakened teacher job protections and enhanced how performance ratings from teacher evaluation systems could be used for high-stakes decisions. In five states, new laws eliminated the ability of new teachers to earn tenure (all between 2011 and 2014). Among the states that maintained tenure protections, twelve increased the number of probationary years during which new teachers can be dismissed without cause (all between 2010 and 2015). Six states restricted or eliminated mandatory collective bargaining for teachers, which limited unions' ability to negotiate over evaluation systems (three between 2003 and 2005 and three in 2011). Finally, four states passed "right to work" laws that eliminated mandatory union dues, creating the possibility that unions would have less funding and lower membership to advocate for things like strong job protections (all between 2012 and 2016).

2.3. Other concurrent education policy reforms

Accountability reforms were not the only education policy reform states implemented during this time period. Several of these reforms, such as new licensure tests and alternative pathways into teaching, largely occurred before the push for teacher accountability reforms. The NCLB Act's high-quality teacher provisions prompted many states to adopt new licensure tests, e.g., basic skills tests (seven states between 2003 and 2008 and nine states between 2012 and 2016), pedagogical knowledge tests (nine states between 2003 and 2005 and four states in 2015), and subject content knowledge tests (14 states between 2003 and 2008 and six states between 2015 and 2016). Between 2002 and 2016, the total number of alternative teacher certification programs doubled from 66 to 132. However, most of these alternative programs were established prior to 2011 (e.g., there were 116 programs in 2010).

Eighteen states won RTT grants. These grants provided additional financial resources, but the four-year total of these awards amounted to <1% of most state education budgets. The influence of the RTT grant competition was far broader as 46 states applied across three RTT funding rounds.[1] For instance, RTT grant scoring criteria incentivized states to implement new college and career ready standards. Between 2011 and 2014, 43 states adopted the Common Core State Standards (CCSS). Another common reform was to increase the proportion of teachers' salaries that were withheld to fund pension obligations, which thirty-two states implemented.[2]

3. Conceptual framework

How might teacher accountability reforms affect the supply and quality of prospective teachers? Evidence suggests that individuals who select into teacher preparation programs place a higher premium on job security than other college graduates (Bowen et al., 2015; Lang and Palacios, 2018). Accountability reforms that weaken job security could decrease new labor supply in the absence of offsetting increases in teacher salaries. Reforms could also decrease new labor supply if they make the profession less enjoyable by decreasing teacher autonomy through scripted curricula aligned with high-stakes tests or an increased focus on test preparation (Reback et al., 2014). Even if accountability reforms have no direct effect on job protections or satisfaction, they may still affect new labor supply if they create the perception among potential entrants that teaching is a less secure or enjoyable career (Kraft and Gilmour, 2016).

Here, we provide an intuitive discussion of the potential effects of teacher accountability reforms on the supply and quality of prospective teachers based on a simple Roy (1951) model.[3] Consider a labor market where individuals choose between a career teaching in public schools or an alternative occupation that represents all outside options by choosing the occupation that maximizes their expected earnings. As noted by Nagler et al. (2020) among others, such a model predicts negative selection on ability into teaching if two conditions are met: 1) ability is valued in both sectors[4] and 2) teaching has lower returns to ability. Evidence from several studies suggests these two conditions are likely to hold in the U.S. context (Britton and Propper, 2016; Chingos and West, 2012; Feng and Sass, 2017; Goldhaber et al., 2007; Hoxby and Leigh, 2004; Lang and Palacios, 2018).

The effect of accountability reforms on the supply of new teachers is unambiguously negative in the model. Similar to Angrist and Guryan (2004, 2008), we conceptualize accountability reforms as increasing the relative costs associated with teaching through the monetized costs of declines in job security and job satisfaction in teaching relative to the alternative occupation. This reduces expected earnings in the teaching profession and leads to an intuitive decline in the share of individuals that choose teaching as a career.

The effect of accountability reforms on the quality of prospective teachers is more ambiguous. The effect of these reforms depends on whether the relative costs associated with teaching are the same for everyone or vary with an individual's ability. When the relative costs are the same for everyone, individuals on the margin between teaching and the alternative occupation are the highest ability prospective teachers. Consequently, when costs increase and supply declines, the average ability of individuals choosing a career in teaching also declines. On the other hand, there is reason to believe *relative* costs decline with ability. For example, passing licensure tests is likely costlier for low-ability individuals. Similarly, reductions in perceived job security or stress associated with high-stakes teacher evaluations may be less salient for higher-ability individuals. If relative costs decline with ability, there is both a high- and low-ability individual on the margin between teaching and the alternative occupation. As a result, when relative costs increase, both the share of high- and low-ability individuals that choose teaching as a career decline, leading to an ambiguous change in new teacher quality (See Appendix Fig. B1).

While teacher accountability reforms may increase the perceived costs associated with teaching, merit pay programs based on evaluation ratings could attract more high-ability teachers into the profession. In theory, merit pay could increase the relative return to ability in the teaching sector enough to offset the cost increases associated with accountability reforms, leading to an increase in the quality of prospective teachers. However, only 20% of states designed some type of bonus or merit pay system (Steinberg and Donaldson, 2016). Even among districts that won federal grants to design and fund merit pay systems, these programs were often poorly understood by teachers, with relatively small bonuses that were awarded to a majority of teachers (Chiang et al., 2017). Very few of these programs were sustained over time (Dee and Wyckoff, 2015). Thus, it remains unclear how accountability might affect

[1] Private foundations and philanthropic organizations such as the Gates and Broad Foundations and the New Schools Venture Fund also invested millions of dollars to support evaluation reforms across the country.

[2] We describe our data sources and coding procedures for these education reforms in Appendix A and provide a complete list of reform dates for each state in Appendix Table A1.

[3] See Appendix B for a more formal analysis.

[4] More specifically the type of ability that is valued in both sectors is positively correlated and the correlation is sufficiently strong to induce migration across sectors.

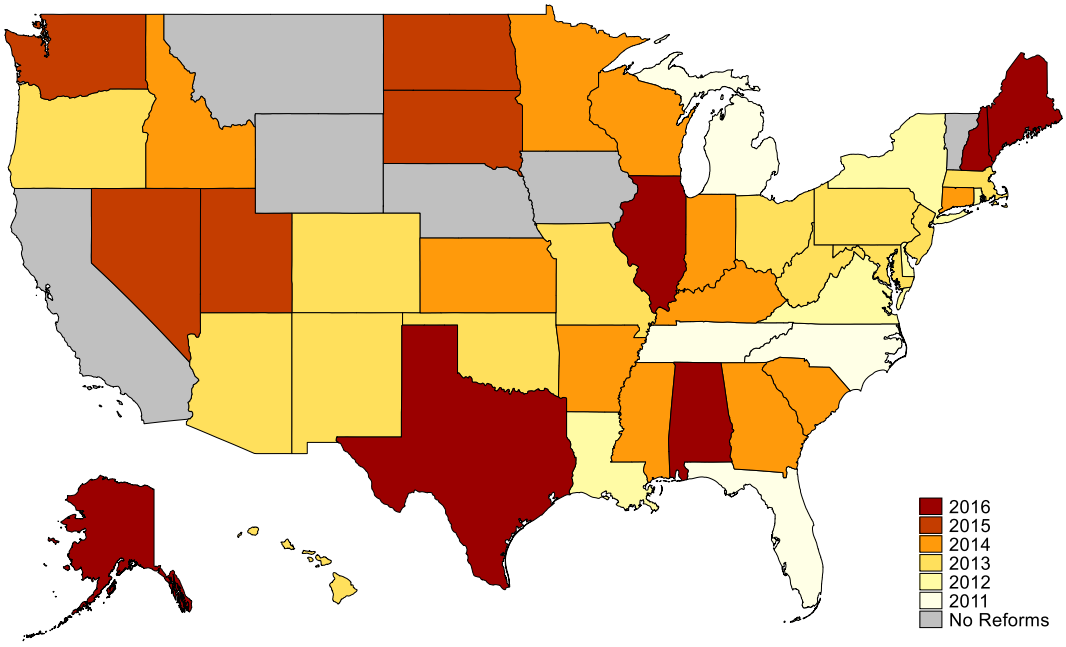

Fig. 1. The timing of statewide implementation of teacher evaluation reforms. Notes: Years represent the fall of the academic year in which the new systems were first fully implemented statewide.

the quality of prospective teachers. Our empirical results help to shed light on this question.

4. Data

We conduct our analyses using an original state-by-year panel from 2002 through 2016 that combines measures from a range of datasets maintained by the U.S. Department of Education, the Bureau of Labor Statistics (BLS), the U.S. Census Bureau, and the National Center for Education Statistics (NCES). We complement these data with repeated waves of the Schools and Staffing Survey (SASS) and the National Teacher and Principal Survey (NTPS). The SASS and NTPS are nationally representative surveys of U.S. schools and teachers conducted by NCES every four years. Critical for our analyses, NCES has maintained a large set of consistent items across administrations on both the Teacher Questionnaire and the School Questionnaire. We use data from the 2003–04, 2007–08, and 2011–12 SASS and 2015–16 NTPS to construct a four-period panel dataset covering the relevant period for evaluation reforms.

4.1. Evaluation reforms

We draw upon two systematic reviews of teacher evaluation reforms to create two binary measures that reflect the timing of evaluation reforms across states (Steinberg and Donaldson, 2016; NCTQ, 2016). We code our preferred measure, *Implement Evaluation,* as one in the fall of the academic year in which the new evaluation system was implemented statewide for the first time. This serves to break much of the collinearity with other education initiatives passed in the same legislative sessions as evaluation reforms. As shown in Fig. 1, states rolled out their new evaluation systems across several years allowing us to jointly model evaluation effects and control for underlying trends in outcomes over time. We test the robustness of our results using an alternative measure, *Passed Evaluation*, which we code as one in the calendar year in which evaluation reforms were passed by state legislatures.

4.2. Accountability reform intensity

Given the related nature of the different accountability reforms passed between 2011 and 2016, we construct a composite measure of these reforms to capture the intensity of reform efforts in a state. Our measure is a simple count measure ranging between 0 and 5 that takes on an incremental value of 1 for each of the five teacher accountability reform measures of the period: adopting high-stakes teacher evaluation systems, eliminating or weakening tenure, increasing the length of the probationary period, eliminating or restricting the scope of mandatory collective bargaining, and eliminating mandatory union dues. We also create an indicator for having exactly one accountability reform versus none and an indicator for having two or more reforms versus none, which we use to explore the dose-response relationship.

4.3. Teacher supply

Ideally, we would measure the supply of new teachers as the total number of candidates that applied for K-12 public school teaching positions for the first time. Although such a measure is unavailable at the national level, data collected by the U.S. ED on the number of initial public-school teacher licenses granted by states each year serves as an advantageous proxy. Any immediate effects of accountability reforms would likely be on the margin of whether teacher preparation program graduates decide to teach in public schools, while delayed effects would likely capture impacts on students' choice to enroll in a teacher preparation program as well.

These data collected under Title II requirements capture all new teachers eligible to work in publicly-funded schools (traditional or charter), regardless of certification pathway or licensure type, and link teacher candidates to their state of intended employment. The range of licenses include professional certifications granted to graduates of traditional preparation programs, initial certifications granted to graduates of alternative pathway programs, and temporary teaching licenses such as emergency, probationary, or intern credentials. Teachers who enter the profession via alternative pathways such as Teach for America are required to have some type of temporary license to teach while they complete the requirements necessary to obtain a provisional teaching license. This measure is also largely robust to the rapid expansion of the charter school sector during our panel period. The vast majority of states require teachers who work in charter schools to obtain a state license.[5] Although licensure reciprocity agreements exist between some states, individuals seeking to teach in a new state must still

[5] Of the 42 states with charter school laws in 2016, 28 states required that teachers obtain state licenses, while 11 states required the majority of teachers to be licensed (around 50 to 75% of teachers). Only three states (Arizona, Alabama [which only allowed charters in 2015], and Louisiana) had no licensure requirement for charter school teachers (Education Commission on the States, 2016).

acquire a state-specific initial teaching license and will be included in our data. Relative to pre-recession levels in 2007, the number of new teaching licenses issued nationally has declined by 23.4%.

4.4. School hiring

We leverage data from the SASS/NTPS School Questionnaire to construct two measures related to schools' experiences filling vacant teaching positions. Schools report on the difficulty of filling vacancies across a range of teaching fields on a four-point Likert scale (*Easy*, *Somewhat Difficult*, *Very Difficult*, *Could not Fill the Vacancy*). We pool responses across 12 subject areas and construct an indicator of whether there were *any* unfilled vacancies at the time the survey was completed (typically in October). This measure of schools' ability to find qualified candidates complements our analyses of teacher supply and quality. It provides a direct measure of supply, in terms of applicant volume, and potentially captures a gradient of teacher quality to the degree that schools screen candidates based on some minimum threshold. Prior to the Great Recession, approximately 4% of schools had at least one unfilled teaching vacancy. In 2011, the proportion of unfilled teaching vacancies dropped by half to 2% but then rose swiftly to 7.1% by 2015.

4.5. Teacher quality

The SASS/NTPS Teacher Questionnaire captures information on the institutions where teachers attended college as well as the type of licensure they hold. These data allow us to merge Barron's rankings of the selectivity of each institutions' admissions process in 2014 as a proxy measure for teacher quality. Barron's rank is measured on a six-point scale ranging from *Non-Competitive* to *Most Competitive*.[6] Research documents a positive, albeit weak, relationship between Barron's rankings and teacher outcomes including pedagogical content knowledge tests, hiring interviews, and lesson demonstrations (Jacob et al., 2016), performance evaluations (Rockoff et al., 2011; Jacob et al., 2016), and value-added to student achievement (Clotfelter et al., 2006; Boyd et al., 2008). At the same time, other studies find no significant relationship between undergraduate selectivity and teachers' contributions to student achievement, conditional on a large set of time-varying teacher characteristics (Clotfelter et al., 2007; Chingos and Peterson, 2011).

To allow the effects of evaluation reforms on new teacher quality to be heterogeneous and non-monotonic, we apply a flexible, non-parametric approach (i.e., a "distribution regression") for examining effects on Barron's rank (Chernozhukov et al., 2013). We construct a set of Barron's ranking indicators where each indicator captures students who attended an undergraduate institution at a given level of competitiveness or higher. We then use those indicators to estimate the entire conditional distribution of teacher quality. We also present results based on a simple binary model where the outcome equals one for teachers that attended undergraduate institutions ranked in the top four categories of *Competitive or higher*.

We construct a second measure of teacher quality based on teachers' certification type. Previous research has found that compared to traditionally certified teachers, teachers with emergency or temporary licenses are less effective at raising student achievement, improve less as they gain experience on the job, and are more likely to leave the profession (Boyd et al., 2006; Clotfelter et al., 2007; Clotfelter et al., 2010). We code this indicator as taking a value of one if a teacher reported holding a "certificate issued to persons who must complete a certification program in order to continue teaching (often called a waiver or emergency certificate)" or if they do not hold a teaching certificate of any type.

4.6. State controls

In our preferred models, we include a parsimonious set of plausibly exogenous control variables to capture state-specific economic conditions and population characteristics. Controls for economic conditions include annual unemployment rates, real average hourly wages in the private sector, and real state tax revenue per capita. Controls for state population characteristics include the percentage of 5-to-17-year-olds that are Black, Hispanic, or White, as well as living in families at or below the federal poverty line.[7]

4.7. Policy controls

We also include a vector of indicators for the concurrent education policy reforms described above. These include indicators for whether a state has won an RTT grant; implemented Common Core state standards; or required teacher candidates to take a basic skills licensure test, a content area licensure test, or a pedagogical content knowledge test. We also include continuous measures for the overall passing rates for licensure tests in a state; the number of alternative certification programs; and the percent of teachers' salary withheld for pension funds.

4.8. School controls

Across all analyses using data from the SASS/NTPS, we pair our set of time-varying state-level controls with controls for school characteristics to account for any random variation in the composition of sampled schools in each wave of the data. These controls include the percent of students that are Black or Hispanic, eligible for free- or reduced-price lunch (FRPL), and have an Individualized Education Plan (IEP), as well as school size, urbanicity, average daily attendance, and type. In Table 1, we provide descriptive statistics for the primary variables used in our analysis to anchor the magnitude of our estimates.

5. Empirical framework

We estimate the effects of teacher evaluation reforms on the decision of individuals to enter the teacher labor market using a difference-in-differences framework. Our identification strategy compares changes within treated states over time to other non-treated states in the same geographic regions. Importantly, the differential timing of accountability reforms across treated states allows us to remove any regional trends in teacher labor supply and demand that might confound our estimates.

We begin by flexibly modeling any anticipatory effects or time-varying treatment effects with a non-parametric event-study specification:

$$Y_{st} = \sum_{r=-7}^{3} 1(t = t_s^* + r)\beta_r + X_{st}\theta + \pi_s + \gamma_{g(s)t} + \varepsilon_{st}, \quad (1)$$

where Y_{st} is an outcome of interest for state s in year t, X_{st} is a vector of time-varying state covariates, π_s and $\gamma_{g(s)t}$ are state and region-by-year fixed effects, respectively, that account for fixed differences in teacher labor markets across states and regional labor market shocks across time, and ε_{st} is a random disturbance term.[8] The term $1(t = t_s^* + r)$ represents a set of indicators for the years pre- and post-policy reform, with t_s^* denoting the year in which state s implemented an evaluation reform and $r \in [-7, 3]$.

The coefficients of primary interest in (1) are the β_r's, which represent the effect of evaluation reforms on our outcomes of interest r years before or after a reform. We measure these effects relative to the year just prior to a reform ($r = -1$) and censor r at -7 and 3 so that β_{-7} and β_3

[6] We include institutions in the Special category as Non-Competitive as they are largely trade schools.

[7] We construct data on average hourly wages in the private sector using annual data from the Current Population Survey (CPS). Data on state unemployment rates are from the BLS. Poverty and racial/ethnic data for the percentage of 5-to-17-year-olds and state tax revenue data come from the U.S. Census Bureau. We link school data to our panel using the spring year of the academic year (e.g., 2015 for AY 2014–15). Hourly wages are deflated to real 2014 dollars using the consumer price index.

[8] We specify $\gamma_{g(s)t}$ using the eight U.S. regions identified by the Bureau of Economic Analysis.

Table 1
Descriptive Statistics.

	Obs	Mean	Std. Dev.	Date Range	Data Source
State Measures					
Teacher Licensures^	750	14.97	5.80	2002-2016	Title II
Program Completers from Graduate TPPs	750	6.31	3.59	2002-2016	IPEDS
Unemployment Rate	750	5.99	1.99	2002-2016	BLS
Average Hourly Wages in the Private Sector	750	27.57	3.50	2002-2016	CPS
Tax Revenue per Capita	750	2.82	1.01	2002-2016	Census
% of 5-17 year olds that are African-American	750	13.07	11.15	2002-2016	Census
% of 5-17 year olds that are Hispanic	750	14.02	12.67	2002-2016	Census
% of 5-17 year olds that are White	750	65.97	17.49	2002-2016	Census
% of 5-17 year olds Below Poverty Line	750	16.89	4.98	2002-2016	Census
Full-time Novice New Teacher Hires^	200	6.94	3.05	2003, '07, '11, '15	SASS/NTPS
Log Public School Total Enrollment	750	13.30	1.02	2002-2016	NCES
Log Private School Total Enrollment	400	10.87	1.20	2002-2016, biannually	NCES
Pupil Teacher Ratio	750	15.38	2.64	2002-2016	NCES
Certified Public Account Exam Takers^	550	2.61	3.41	2006-2016	AICPA
Average Public Teacher Salary (NCES)	750	54,999	8,054	2002-2016	NCES
Adjusted Average Public Teacher Salary (ACS)	600	46,625	6,577	2005-2016	ACS
School Measures					
At least one unfilled vacancy in a school	28,610	0.04	0.20	2003, '07, '11, '15	SASS/NTPS
Teacher Measures					
Worry About Job Security	6,460	2.30	0.99	2003, '07, '11, '15	SASS/NTPS
Job Satisfaction	6,460	3.45	0.75	2003, '07, '11, '15	SASS/NTPS
Teacher Cooperation	6,460	3.26	0.78	2003, '07, '11, '15	SASS/NTPS
Control over Selecting Instructional Materials	6,460	2.44	1.05	2003, '07, '11, '15	SASS/NTPS
Control over Content, Topics, Skills Taught	6,460	2.75	1.03	2003, '07, '11, '15	SASS/NTPS
Control over Teaching Techniques	6,460	3.58	0.64	2003, '07, '11, '15	SASS/NTPS
Competitive Undergraduate Institution or Higher	5,800	0.85	0.36	2003, '07, '11, '15	SASS/NTPS
Emergency or Temporary Certification	5,800	0.16	0.36	2003, '07, '11, '15	SASS/NTPS

Notes: Teacher characteristics are weighted using appropriate probability weights from the SASS dataset. Per pupil expenditures, average hour wages in the private sector and state tax revenue per capita are reported in 2014 real dollars. Missing values are imputed for Full-time Novice New Teacher Hires using linear interpolation. SOURCE: Title II = U.S. Department of Education Title II Data System, IPEDS = Integrated Postsecondary Education Data System, BLS = Bureau of Labor Statistics, CPS = Current Population Survey, Census = U.S. Census Bureau, SASS = Schools and Staffing Survey, NTPS = National Teacher and Principal Survey, NCES = National Center for Education Statistics, AICPA = American Institute of Certified Public Accountants, ACS = American Community Survey. Sample sizes are rounded to the nearest tens place per Institute for Education Sciences (IES) disclosure guidelines.
^ scaled per 10,000 18-65 year olds.

represent the average effect of reforms on our outcomes of interest 7 or more years prior to a reform and 3 or more years after a reform, respectively.

We complement the event-study specification with a standard difference-in-differences (DD) model to increase our precision by pooling estimates across post-reform years:

$$Y_{st} = 1(t \geq t_s^*)\beta_1 + X_{st}\theta + \pi_s + \gamma_{g(s)t} + \mu_{st}, \quad (2)$$

where, $1(t \geq t_s^*)$ represents an indicator variable that equals unity in all years post-policy adoption, μ_{st} is a random disturbance term and all other variables are as defined in (1). The coefficient of primary interest in (2) is β_1, which is the difference-in-differences estimate of the effect of a given policy reform averaged across the post-period years in our panel.

Finally, we relax the assumption of time-invariant treatment effects imposed by the standard DD model in (2) to more formally test for any incremental effects and differential pre-trends in outcomes (Goodman-Bacon, 2018). Specifically, we follow Lafortune et al. (2018) and add two linear time trends as follows:

$$Y_{st} = 1(t \geq t_s^*)\beta_1 + 1(t \geq t_s^*)(t - (t_s^* - 1))\beta_2 + (t - (t_s^* - 1))\beta_3 + X_{st}\theta + \pi_s + \gamma_{g(s)t} + \upsilon_{st}, \quad (3)$$

where $(t - (t_s^* - 1))$ is a linear trend centered at zero on the year before a state implemented an evaluation reform, υ_{st} is a random disturbance term and all other variables are as defined in (2).[9] The interaction term in (3), $1(t \geq t_s^*)(t - (t_s^* - 1))$ allows for the relative time trends among treated states to differ pre- and post-reform. β_1 captures the immediate response of the policy change on our outcome of interest, while β_2 captures any deviation from the linear trend in labor supply in the post-reform period among treated states. β_3 tests for any differential linear trends in the pre-reform period among states that adopted teacher accountability reforms relative to those that did not. We report standard errors clustered at the state level in all tables.

We account for differences in the size of state labor markets by scaling our measures of new teacher labor supply relative to the size of the pool of potential new entrants in the teaching profession (i.e. per 10,000 individuals aged 18-to-65 in a given state and year).[10] When the outcome of interest is this scaled measure, we also weight our models by the number of individuals aged 18-to-65 per 10,000.[11] We apply the appropriate SASS/NTPS probability sampling weights for our analyses of school hiring and teacher quality outcomes. These approaches allow us to recover nationally representative estimates of the effect of teacher evaluation reforms, improve the precision of our estimates, and account for the endogenous sampling framework used in the SASS/NTPS (Solon et al., 2015).

Our overall DD framework relies on two key assumptions: 1) that comparison states provide a valid counterfactual for the trends in treated states, and 2) that there are no unobserved factors that are correlated with both our outcomes of interest and the timing of teacher evaluation reforms across states. We test the first assumption visually via our non-parametric event study specifications and formally via differential linear pre-trends in our DD model. To examine the validity of

[9] This allows the linear time trend to take on a value of 1 in the first year of the reform ($r = 0$) making the treatment effect a linear combination of the intercept shift β_1 and the change in slope β_2. Centering on the year of evaluation reforms produces identical results but causes the joint effect of β_1 and β_2 to load entirely on β_1.

[10] In Appendix Table A7 and A9, we show that our results are robust to scaling by the number of 22–25 year olds per 1,000.

[11] Weighting serves to increase the precision of our estimates because the number of licensures granted varies considerably within states over time (ICC = 0.35).

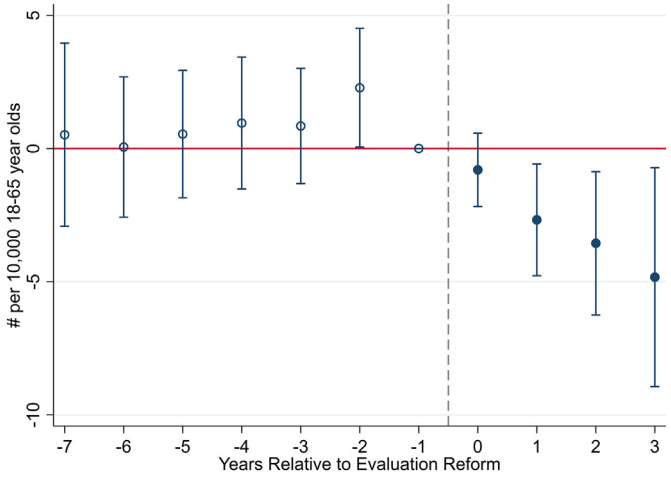

Fig. 2. Event study depicting effect of evaluation reforms on the number of new teaching licenses. Notes: Point estimates for years pre- (hollow dot) and post-reform (solid dot) and corresponding 95% confidence intervals are from models without controls reported in Appendix Table A2.

the second assumption, we run a series of auxiliary regressions where we regress our exogenous demographic and economic state controls as well as endogenous measures for new teacher labor demand on our treatment indicator with and without region-by-year and state fixed effects. It is reassuring that these auxiliary regressions present little evidence that accountability reforms affected the demand for new teachers, or that these reforms coincided with changing student demographics or broader negative economic shocks that decreased labor demand across state labor markets.

6. Results

6.1. Effects on new teacher supply

We begin by presenting estimates from a simple event-study model of the effect of evaluation reforms on the number of new teaching licenses. As shown in Fig. 2, the parameter estimates with associated 95% confidence intervals for the years pre-reform (hollow dot) and post-reform (solid dot) are strongly suggestive of a negative effect of evaluation reforms on the supply of new teachers.[12] The number of teacher licenses granted declines steadily, starting in the year that high-stakes evaluation reforms were implemented statewide. These individual point estimates become statistically significant in the second year and suggest that the effects of the reforms increase over time. Importantly, there is little evidence that the supply of new teachers was trending downward prior to the implementation of evaluation reforms.

Results from our standard DD model confirm the effect of high-stakes evaluation reforms. In Table 2, we estimate that high-stakes evaluation reforms reduced licenses granted in a state by 2.69 per 10,000 18-to-65-year-olds, on average, in our baseline specification without controls. This represents a 17% reduction in the average number of licenses granted in the post-policy reform years among treated states, relative to the pre-reform state mean. As shown in column 2, this estimate is quite robust to the inclusion of controls for state-specific economic conditions, student population characteristics, and other education policy reforms.

Estimates based on our parametric DD specification that includes linear trends also confirm the patterns suggested by our event-study analyses. Focusing on results from our model that includes controls in column 4 of Table 2, our estimates reveal a significant downward linear trend in supply among treated states in the post-policy reform years of −1.47 licenses per 10,000 18-to-65-year-olds. Consistent with the visual evidence in the event study, we also fail to reject the null hypothesis of no differential pre-reform trends for states that adopted evaluation reforms.

In columns 5 and 6, we add contemporaneous accountability reform measures to the model. In our model with controls, the estimates for these complementary reforms are all negatively signed as we hypothesized, but the limited variation in these measures produces large corresponding confidence intervals. When we model these related reforms using our single continuous count measure we find strong evidence that the aggregate intensity of accountability reforms affected the overall magnitude of decline in new teacher labor supply. As shown in columns 7 and 8, our linear term capturing the number of accountability reforms adopted is negative and significant, suggesting an approximate 10% decline for each additional accountability reform passed in a state. Estimates using our pair of accountability indicators shown in columns 9 and 10 suggest a more modest dose-response relationship.

6.2. Effects on teacher hiring

We next examine whether, in addition to reducing new teacher labor supply, evaluation reforms also made the process of filling teaching vacancies more challenging for schools. We present results from a simple event-study model in Fig. 3 (see Appendix Table A2 column 3). The pattern of results clearly suggests that evaluation reforms increased the difficulty of filling vacant positions with a growing effect over time. We find an isolated negative point estimate in the pre-period at $t = -3$, possibly the result of the unbalanced sample given the four-year gaps between each SASS/NPTS.

In column 2 of Table 3, we estimate that evaluation reforms increased the probability that a school had at least one unfilled vacancy by 2.6 percentage points in our model that includes state, policy and school controls. Our linear DD estimates in column 3 and 4 suggest these effects were likely both immediate and increasing over time. While there is some evidence of a very small positive and statistically significant pre-trend in the probability of unfilled vacancies (0.002 SD), it is reassuring that even when allowing for this pre-trend we continue to find effects of similar magnitude in our linear DD model as in our standard DD model.

As shown in columns 5 and 6, adding our full set of concurrent accountability measures again has little effect on our estimate of the impact of evaluation reforms. Here, the accountability measures are all positively signed, as we would expect, and relatively small in magnitude with the exception of weakening collective bargaining, which we estimate decreased the likelihood of a vacant position by 1.4 percentage points. We caution against placing too much weight on this estimate given the difficulty of interpreting it conditional on all the other related accountability reforms. Instead, we prefer evidence from the linear accountability intensity measure which suggests that the probability of a school having at least one unfilled vacancy went up by approximately 1 percentage point for each accountability reform a state adopted (columns 7 and 8). Estimates using our pair of accountability indicators also suggest a strong dose-response relationship where the coefficient on 2 or more accountability reforms is twice the magnitude of the coefficient for 1 or more accountability reforms (columns 9 and 10).

In Table 4, we test for heterogeneous effects of evaluation reforms by several proxy measures for hard-to-staff schools. Specifically, we present results from a set of standard DD models where we interact the implement evaluation indicator with school-level measures of: 1) whether the school is located in an urban area; 2) the percent of FRPL students; 3) the percent of students who are Black and Hispanic; and 4) the percent of students with an IEP plan. We find consistent evidence across all four measures that evaluation reforms had a larger impact on the ability of hard-to-staff schools to fill vacant positions. We estimate the effect of evaluation reforms on the probability of having at least one unfilled teaching position was more than twice as large for urban schools as

[12] We report point estimates and standard errors from the event study specification in column 1 of Appendix Table A2.

Table 2
The Effect of Teacher Accountability Reforms on the Number of New Teaching Licenses.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Implement Evaluation	−2.686**	−2.996***	−1.061	−1.256	−2.393**	−2.599***				
	(0.836)	(0.722)	(0.982)	(0.907)	(0.840)	(0.729)				
Implement Evaluation * Trend			−1.308*	−1.468**						
			(0.527)	(0.456)						
Trend			0.231	0.188						
			(0.212)	(0.177)						
Eliminate Tenure					−1.482	−0.510				
					(1.803)	(1.693)				
Increase Probationary Period					−1.003	−1.972				
					(1.306)	(1.361)				
Weaken Collective Bargaining					0.617	−0.533				
					(1.951)	(1.663)				
Eliminate Mandatory Union Dues					−0.687	0.735				
					(1.970)	(2.815)				
Total Accountability Reforms							−1.175*	−1.494**		
							(0.454)	(0.448)		
1 Accountability Reform									−2.605**	−2.880**
									(0.761)	(0.835)
2 or more Accountability Reforms									−2.515+	−3.142*
									(1.286)	(1.249)
% change relative to state mean	−17%	−18%			−15%	−16%	−7%	−9%		
State Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Policy Controls	No	Yes	No	Yes	No	Yes	NO	Yes	No	Yes
n	750	750	750	750	750	750	750	750	750	750

Notes: +p<.10, *p<.05, **p<.01, ***P<.001. Cells report estimates and associated standard errors clustered at the state level in parentheses. Time-varying state controls include real average hourly wages in the private sector, unemployment rates, real state tax revenue per capita, the fraction of the population ages 5 to 17 that are White, Black, Hispanic, and living below the federal poverty line. Policy controls include indicators for having won a Race to the Top grant; implementing Common Core state standards; requiring teacher candidates to take a basic skills licensure test, a content area licensure test, and a pedagogical content knowledge test; the number of alternative certification programs; and the percent of teachers' salary withheld for pension funds. All models include state and region-by-year fixed effects. All models are weighted by the number of 18-65 year olds per 10,000 in a state. Sample sizes are rounded to the nearest tens place per IES disclosure guidelines. SOURCE: See Table 1.

for non-urban schools (4.4 vs. 2.0 percentage points). As shown in Fig. 4, similar patterns exist for schools with higher percentages of students that are eligible for FRPL, that are Black and Hispanic, and that have IEPs, all of which are scaled so that a one-unit change is a ten percentage point change.

6.3. Effects on new teacher quality

How did teacher evaluation reforms affect the quality of newly hired novice teachers in public schools? Given potential heterogeneous and non-monotonic effects of evaluation on teacher quality, we first estimate models where our outcomes are a set of indicators capturing whether a teacher attended an undergraduate institution of a given rank or higher. As show in Fig. 5 and Appendix Table A3, the pattern of results across the rankings is strongly suggestive of a rightward shift in the quality of the marginal low-quality teacher, with little change in the quality of the marginal high-quality teacher. These findings are consistent with a simple Roy model of occupational choice where costs vary by ability. They suggest that evaluation reforms shifted the supply of new teachers upward along the quality distribution by primarily reducing the probability that teachers graduated from non-competitive institutions and increasing the probability they graduated from competitive institutions.

In our preferred model in Table 5 Panel A, we find that evaluation reforms increased the probability a teacher graduated from an undergraduate institution ranked competitive or higher by 8.1 percentage points. Estimates from linear DD models suggest these impacts were largely immediate and continued to increase in the post-reform years. In Panel B of Table 5 we present estimates of the effect of evaluation reforms on the probability that a newly hired novice teacher held an emergency certification. We find no effect of evaluation reforms across both standard and linear DD models suggesting that the reforms had little effect on the number of new teachers who lacked more formal state licenses.[13]

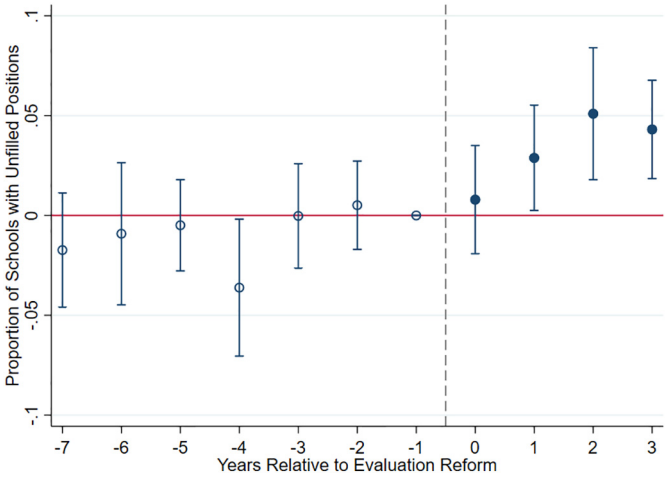

Fig. 3. Event study depicting effect of evaluation reforms on school hiring: schools with at least one unfilled teaching position. Notes: Point estimates for years pre- (hollow dot) and post-reform (solid dot) and corresponding 95% confidence intervals are from models without controls reported in Appendix Table A2.

6.4. Mechanisms

We leverage a rich set of self-reported survey questions across waves of the SASS/NTPS to explore the ways in which evaluation reforms might have increased the relative cost of entering the teaching

[13] In supplemental analyses, we find no evidence that effects on teacher quality differed by school urbanicity or the percentage of students in a school that are eligible for FRPL, that are Black and Hispanic, and that have IEPs.

Table 3
The Effect of Teacher Accountability Reforms on the Probability Schools Have at least One Unfilled Teaching Position.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Implement Evaluation	0.032**	0.026***	0.013	0.010	0.030**	0.026**				
	(0.009)	(0.007)	(0.009)	(0.008)	(0.011)	(0.008)				
Implement Evaluation * Trend			0.008**	0.006*						
			(0.003)	(0.003)						
Trend			0.002	0.002*						
			(0.001)	(0.001)						
Eliminate Tenure					0.009	0.013				
					(0.008)	(0.013)				
Increase Probationary Period					0.007	0.007				
					(0.009)	(0.006)				
Weaken Collective Bargaining					−0.006	−0.016*				
					(0.007)	(0.007)				
Eliminate Mandatory Union Dues					0.015	0.007				
					(0.019)	(0.015)				
Total Accountability Reforms							0.013***	0.009**		
							(0.003)	(0.003)		
1 Accountablitiy Reform									0.020*	0.015*
									(0.009)	(0.007)
2 or more Accountability Reforms									0.030***	0.027***
									(0.008)	(0.007)
School Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
State Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Policy Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
n	750	750	750	750	750	750	750	750	750	750

Notes: +p<.10, *p<.05, **p<.01, ***P<.001. Cells report estimates and associated standard errors clustered at the state level in parentheses. Time-varying state controls and policy controls are the same as those listed in Table 2. School controls include the percent of students that are Black and Hispanic, eligible for free- or reduced-price lunch, and receive services as part of an Individualized Education Plan (IEP) as well as school size, urbanity, and grade level. All models include state and region-by-year fixed effects and use appropriate sampling weights provided by the SASS/NTPS. Sample sizes are rounded to the nearest tens place per IES disclosure guidelines. SOURCE: See Table 1.

profession. These measures include new teachers' responses to statements about job security, job satisfaction, and cooperative effort among teachers on a 4-point Likert scale from *Strongly Disagree* to *Strongly Agree*. We complement these three measures with new teachers' responses on a 4-point Likert scale from *No Control* to *A Great Deal of Control* about how much control they have in their classrooms over: selecting textbook and instructional materials; selecting content, topics, and skills to be taught; and selecting teaching techniques.

We find consistent evidence that evaluation reforms increased the perceived costs of teaching among new entrants into the profession. Fig. 6 displays predicted average marginal effects from ordered logistic regressions based on our standard DD specification with controls (see Appendix Table A4 for results from underlying ordered logistic

Table 4
Differential Effects of Teacher Evaluation Reforms on the Probability Schools Have at least One Unfilled Teaching Position.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Implement Evaluation (Eval)	0.026**	0.020*	0.018*	0.010	0.019+	0.010	0.022*	0.016*
	(0.008)	(0.008)	(0.009)	(0.008)	(0.009)	(0.007)	(0.010)	(0.007)
Urban	0.019**	0.005						
	(0.006)	(0.006)						
Eval * Urban	0.023+	0.024+						
	(0.013)	(0.013)						
Percent FRPL			0.003***	0.001+				
			(0.001)	(0.001)				
Eval * Percent FRPL			0.002+	0.003*				
			(0.001)	(0.001)				
Percent Black and Hispanic					0.005***	0.003***		
					(0.001)	(0.001)		
Eval * Percent Black and Hispanic					0.004*	0.005**		
					(0.002)	(0.002)		
Percent IEP							0.001	−0.002
							(0.001)	(0.001)
Eval * Percent IEP							0.007**	0.008**
							(0.003)	(0.003)
School Controls	No	Yes	No	Yes	No	Yes	No	Yes
State Controls	No	Yes	No	Yes	No	Yes	No	Yes
Policy Controls	No	Yes	No	Yes	No	Yes	No	Yes
n	28,610	28,610	28,610	28,610	28,610	28,610	28,610	28,610

Notes: +p<.10, *p<.05, **p<.01, ***P<.001. Cells report estimates and associated standard errors clustered at the state level in parentheses. Percent FRPL, Black and Hispanic, and IEP measure are scaled so that a one-unit change is equivalent to a ten percentage point change. See Table 2 notes for a complete list of time-varying state and policy controls and Table 3 notes for a complete list of school controls. All models include state and region-by-year fixed effects and use appropriate sampling weights provided by the SASS/NTPS. FRPL = Free or Reduced Price Lunch, IEP = Individualized Education Plan. Sample sizes are rounded to the nearest tens place per IES disclosure guidelines. SOURCE: See Table 1.

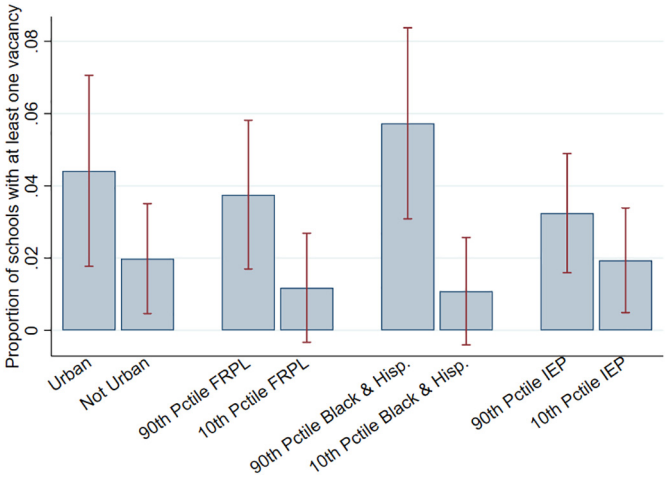

Fig. 4. The heterogeneous effects of evaluation reforms on school hiring: the probability schools had at least one unfilled teaching position by school characteristics. Notes: Bars represent 95% confidence intervals. Estimates based on models that include state, policy, and school controls reported in Table 4.

Table 5
The Effect of Teacher Evaluation Reforms on New Teacher Qualifications.

	(1)	(2)	(3)	(4)
Panel A. Competitive Undergraduate Institution or Higher				
Implement Evaluation	0.046	0.081**	0.010	0.073*
	(0.029)	(0.028)	(0.025)	(0.031)
Implement Evaluation * Trend			0.022+	0.017
			(0.011)	(0.012)
Trend			−0.002	−0.008*
			(0.003)	(0.004)
n	5,800	5,800	5,800	5,800
Panel B. Emergency or No Certification				
Implement Evaluation	−0.008	0.006	−0.009	−0.026
	(0.030)	(0.033)	(0.029)	(0.037)
Implement Evaluation * Trend			−0.000	0.017
			(0.009)	(0.011)
Trend			0.000	0.002
			(0.003)	(0.005)
n	6,460	6,460	6,460	6,460
School Controls	No	Yes	No	Yes
State Controls	No	Yes	No	Yes
Policy Controls	No	Yes	No	Yes

Notes: +p<.10, *p<.05, **p<.01, ***P<.001. Cells report estimates and associated standard errors clustered at the state level in parentheses. See Table 2 notes for a complete list of time-varying state and policy controls and Table 3 notes for a complete list of school controls. All models include state and region-by-year fixed effects and use appropriate sampling weights provided by the SASS/NTPS. Sample sizes are rounded to the nearest tens place per IES disclosure guidelines. SOURCE: See Table 1.

regression models). We find that evaluation reforms increased the number of teachers who *Agree* and *Strongly Agree* that they worry about job security by 8.8 and 7.7 percentage points, respectively. We depict an event study version of these results in Fig. 7 using a binary outcome of whether teachers *Agree* or *Strongly Agree* that they worry about job security. Here, we see no pre-trend but evidence that novice teachers began to worry about job security a year before evaluation reforms were fully implemented. This is suggestive of an anticipatory effect on how much teachers worried about their job security as districts were preparing to fully implement new evaluation systems and teachers were aware of these impending changes.

Evaluation reforms also appear to have reduced teacher satisfaction and autonomy. We find that evaluation reforms resulted in a 14.6 percentage point drop in the likelihood teachers *Strongly Agree* that they are satisfied with being a teacher. Similarly, we find a 12.7 percentage point drop in the proportion of teachers who *Strongly Agree* that there exists a great deal of cooperative effort among teachers. We find a 5.7 percentage point decrease in the probability that new teachers *Strongly Agree* that they have control over the content and skills they teach and an 8.9 percentage point drop in the probability that new teachers *Strongly Agree* that they have control over their teaching techniques. We find no effects on teachers' perceptions about their control over selecting instructional materials. Together, these analyses suggest that evaluation reforms substantially decreased new teachers' perceived job security, job satisfaction, cooperative effort, and control over content and teaching methods.

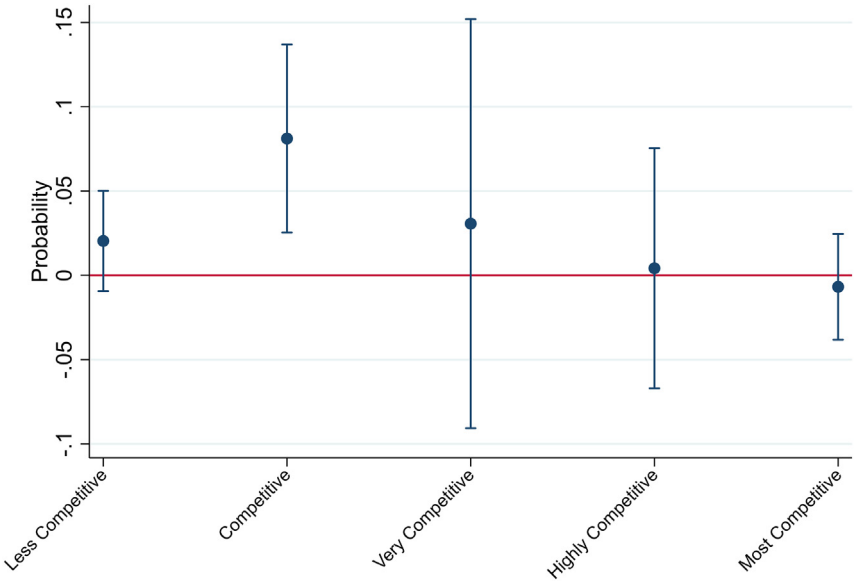

Fig. 5. Point estimates of the effect of evaluation reforms at different points in the distribution of the selectivity of newly hired novice teachers' undergraduate institutions. Notes: Each estimate is associated with a binary outcome measure that divides the Barron's ranking scale into two groups, the ranking level labeled in the figure or higher rankings (which take on a value of one) versus all lower ranking levels (which take on a value of zero).

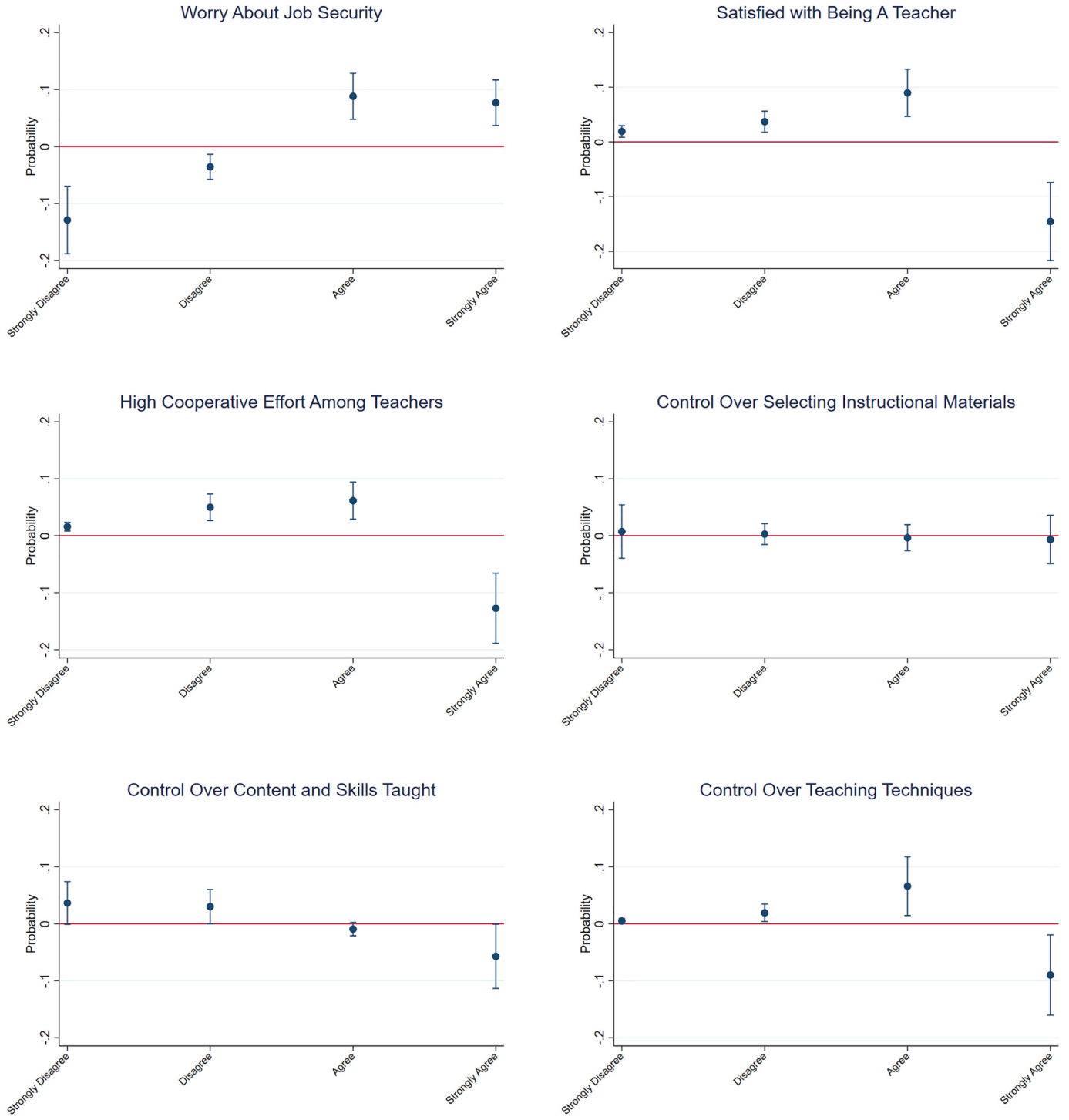

Fig. 6. Predicted marginal effects of evaluation reforms on new teachers' perceived working conditions.

7. Robustness tests

7.1. Alternative modeling and weighting approaches

In Appendix Tables A5 and A6, we show that our results are robust to alternative modeling approaches that include defining treatment as the year teacher evaluation reform laws were passed instead of implemented, replacing region-year fixed effects with year fixed effects, restricting the sample to the 29 states that implemented evaluation reforms but no other accountability reforms, including endogenous controls for teacher demand, and allowing for pre- and post-trends for accountability and other education policy reforms. Estimated effects for licenses and vacancies are statistically and economically significant in all but one model. Estimates of the effect of passing an evaluation reform on vacancies lose significance in a simple DD specification, but we do find evidence of delayed impacts in our more flexible linear DD model. In Appendix Table A7, we confirm that our results for licenses are robust to applying different scaling and weighting factors, dropping weights, and logging rather than scaling new teacher labor supply. Finally, in Fig. A1 we show that our event study results are consistent when we extend the range of bins to $r \in [-10, 4]$.

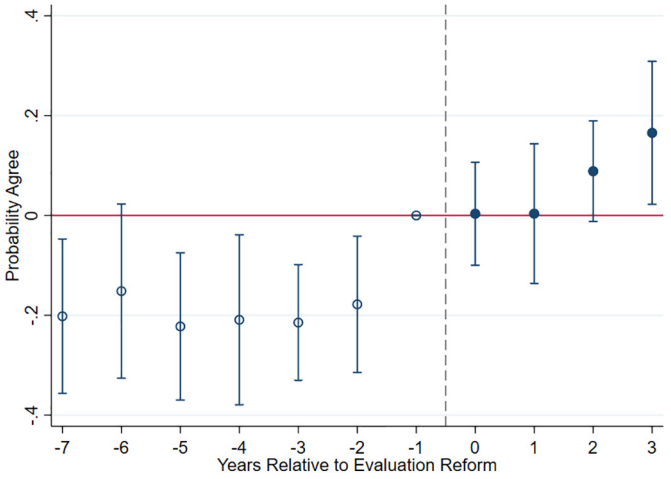

Fig. 7. Event study depicting effect of evaluation reforms on probability teachers agree that they worry about job security. Notes: Point estimates for years pre- (hollow dot) and post-reform (solid dot) and corresponding 95% confidence intervals are from models without controls reported in Appendix Table A2.

7.2. Covariate balance tests

We further explore the degree to which our core DD model sufficiently accounts for any correlation between changing demographics of the student-age population or negative economic shocks and states' implementation of evaluation reforms. As shown in Table 6, in simple bivariate specifications there is a relationship between the timing of evaluation implementation and several of our covariates. However, adding our identifying controls, state and region-by-year fixed effects, substantially reduces these relationships such that none of them represent more than a 3% change relative to the pre-reform state mean and only one is marginally significant. One exception is that with the addition of basic DD fixed effects, evaluation implementation predicts a drop in state tax revenue per capita of approximately 6%. Nevertheless, it is reassuring that results are quite robust to including this control in our models (Table 7).

7.3. Falsification tests and shocks to teacher demand

Given the relationship between our treatment indicator and state tax revenue per capita, we conduct a falsification test to explore whether evaluation reforms appear to affect prospective entrants into a private sector industry that also requires a bachelors' degree and state certification: accounting. Specifically, we estimate models where the outcome is a measure of the supply of accountants, first time Certified Public Accountant (CPA) exam takers in a given state and year. As shown in Panel B columns 7 and 8 of Table 6, we find no evidence of a broader effect on labor supply outside of the teacher sector.

Contemporaneous shocks to teacher demand could also cause prospective teachers to update their expectations about job prospects and lifetime earnings in the teaching sector, causing some to choose not to teach. Testing for exogenous shocks to teacher labor demand is challenging for two reasons: 1) there is no national measure of demand that is independent of supply, such as the number of open positions posted, and 2) it is possible that evaluation reforms may have affected demand through effects on teacher turnover and retirement rates. Our approach is to test for evidence of large demand shocks caused by, or concurrent with, teacher evaluation reforms that might explain our results. We begin with an intuitive but endogenous measure of demand – the number of new public school teachers hired. We then test for effects on proxies of teacher demand that suffer from endogeneity to a lesser degree: public and private school enrollment, and pupil-teacher ratios. As shown in Panel C of Table 6, these falsification tests reveal no evidence to suggest that changes in teacher demand are driving the effects we see on new teacher labor supply.

8. Extensions

8.1. Effects on teacher preparation program completers

We further explore the effect of accountability reforms on the number of graduates from university-based teacher preparation programs

Table 6
Auxiliary Regressions Examining the Effect of Teacher Evaluation Reforms on Teacher Labor Demand and Economic Conditions.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Panel A: Student Demographics							
	Percent of 5 to 17 year-olds that are African American		Percent of 5 to 17 year-olds that are Hispanic		Percent of 5 to 17 year-olds that are white		Percent of 5 to 17 year-olds that are from Low-Income Households	
Implement Evaluation	2.841*	0.375+	−3.957	0.085	1.476	−0.331	2.030**	0.408
	(1.397)	(0.197)	(4.808)	(0.278)	(4.539)	(0.332)	(0.680)	(0.274)
n	750	750	750	750	750	750	750	750
% change relative to state mean	22%	3%	−31%	1%	2%	0%	13%	3%
	Panel B: Economic Conditions							
	Unemployment Rate		Private Hourly Wages		State Tax Revenue Per Capita (real dollars)		Certified Public Accountant Exam Takers (per 10,000 18-65 year olds)	
Implement Evaluation	−0.498+	0.066	2.832***	0.050	−0.035	−0.168**	−0.176	0.011
	(0.269)	(0.172)	(0.650)	(0.188)	(0.178)	(0.061)	(0.192)	(0.108)
n	750	750	750	750	750	750	550	550
% change relative to state mean	−9%	1%	11%	0%	−1%	−6%	−7%	0%
	Panel C: Teacher Labor Demand							
	Novice Public School Teachers (per 10,000 18-65 year olds)		Public School Enrollment (in logs)		Private School Enrollment (in logs)		Public School Pupil Teacher Ratio	
Implement Evaluation	−0.806	0.862	−0.191	−0.004	−0.308	0.003	−0.907	−0.243
	(0.720)	(0.579)	(0.243)	(0.009)	(0.236)	(0.020)	(0.915)	(0.260)
n	200	200	750	750	400	400	750	750
% change relative to state mean	−10%	11%	−1%	0%	−3%	0%	−6%	−2%
State & Region-by-Year FE	No	Yes	No	Yes	No	Yes	No	Yes

Notes: +p<.10, *p<.05, **p<.01, ***P<.001. State and policy controls are omitted from all models. Cells report estimates and associated standard errors clustered at the state level in parentheses. All models are weighted by the number of 18-65 year olds per 10,000 in a state. Sample sizes are rounded to the nearest tens place per IES disclosure guidelines. SOURCE: See Table 1.

Table 7
The Effect of Teacher Accountability Reforms on the Number of Graduates from M.A. Granting Teacher Preparation Programs.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Implement Evaluation	−0.555	−0.674*	−0.337	−0.324	−0.402	−0.540+				
	(0.472)	(0.322)	(0.234)	(0.215)	(0.489)	(0.316)				
Implement Evaluation * Trend			−0.263	−0.287						
			(0.204)	(0.183)						
Trend			0.114	0.031						
			(0.106)	(0.083)						
Eliminate Tenure					−0.470	−0.990				
					(0.537)	(0.701)				
Increase Probationary Period					0.195	0.352				
					(0.406)	(0.390)				
Eliminate Mandatory Union Dues					−1.114	−0.669				
					(0.667)	(0.579)				
Weaken Collective Bargaining					0.757	−0.089				
					(0.629)	(0.556)				
Total Accountability Reforms							−0.271	−0.327+		
							(0.183)	(0.171)		
1 Accountability Reform									−0.526	−0.606
									(0.452)	(0.378)
2 or more Accountability Reforms									−0.588	−0.936+
									(0.463)	(0.476)
% change relative to state mean	−9%	−10%			−6%	−4%	−4%	−5%		
State Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Policy Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
n	750	750	750	750	750	750	750	750	750	750

Notes: +p<.10, *p<.05, **p<.01, ***P<.001. Cells report estimates and associated standard errors clustered at the state level in parentheses. Time-varying state controls include real average hourly wages in the private sector, unemployment rates, real state tax revenue per capita, the fraction of the population ages 5 to 17 that are White, Black, Hispanic, and living below the federal poverty line. Policy controls include indicators for having won a Race to the Top grant; implementing Common Core state standards; requiring teacher candidates to take a basic skills licensure test, a content area licensure test, and a pedagogical content knowledge test; the number of alternative certification programs; and the percent of teachers' salary withheld for pension funds. All models include state and region-by-year fixed effects. All models are weighted by the number of 18-65 year olds per 10,000 in a state. Sample sizes are rounded to the nearest tens place per IES disclosure guidelines. SOURCE: See Table 1.

(TPP) using data from the Integrated Postsecondary Education Data System (IPEDS).[14] We consider these results as exploratory given several limitations of the IPEDS data for capturing state-specific new teacher labor supply.[15] We focus our primary analyses on degree completers from graduate programs because we expect enrollment and persistence in these predominantly one-year programs to be more immediately responsive to evaluation reforms and to be a more direct measure of supply. Slightly more than half of all TPP completers are from graduate degree programs.

Our results from a simple event study shown in Fig. 8 (and Appendix Table A2) closely mirror the pattern of effects we find on licensure and vacancies. We see a relatively flat pre-trend and a steady decline in completers from graduate TPPs post reform although the individual point estimates lack precision. In our standard DD model with covariates, we estimate that evaluation reforms reduced the number of completers by 10%. These results are largely unchanged when we include other contemporaneous accountability reform measures. Our estimates of the effect of accountability reform intensity is negative and marginally significant, suggesting that for every additional accountability reform a state adopted, the number of graduates from Masters' granting TPPs declined by 5%. Estimates where we model accountability as two indicators also confirm this positive dose-response relationship (columns 9 and 10).[16] Models using alternative specifications and weights in Appendix Tables A8 and A9 produce results of consistent magnitude.

8.2. Effects on teacher compensation

We estimate the effect of evaluation reforms on two different measures of teacher wages. The first is real average public school teacher wages calculated using district reported total FTE instructional staff salaries collected by NCES. The second is adjusted average total real wages for public school teachers estimated using the 2005–2016 Public Use Microdata Sample (PUMS) of the American Community Survey (ACS). We residualize ACS wages using a quadratic function of age and fixed effects for highest degree earned to account for differences in average wages caused by changes in the distribution of experience and educational attainment among public school teachers rather than changes in the underlying salary schedules. We find no evidence of any compensating differentials that might explain or offset the increased occupational costs caused by teacher evaluation reforms (See Appendix Table A10).

8.3. The effects of other education reform policy reforms

In Table 8, we present results from standard DD models with controls where we estimate the effect of accountability reforms as well as other education policy reforms both separately and simultaneously. When entered alone, our point estimate of the effect of eliminating tenure suggests an equally large negative impact on new teacher supply as the effect of evaluation reforms, but the estimate is very noisy. In joint

[14] See Appendix C for a description of the Classification of Instruction Program (CIP) codes we used to identify teacher preparation program completers.

[15] First, IPEDS does not include graduates of alternative non-degree granting TPPs that are not affiliated with a college or university such as Teach for America and TNTP Teaching Fellows programs in some states. Second, program completers may intend to teach in private schools or work in a non-teaching position. Third, program completers may intend to teach in a state other than the one in which they completed their degree. Fourth, any effects would likely be delayed given students apply to one-year graduate TPPs roughly two years prior to entering the teacher labor market and up to four years prior for bachelor's programs.

[16] In results not shown, we find little evidence that teacher accountability reforms affected the number of TPP graduates from undergraduate programs.

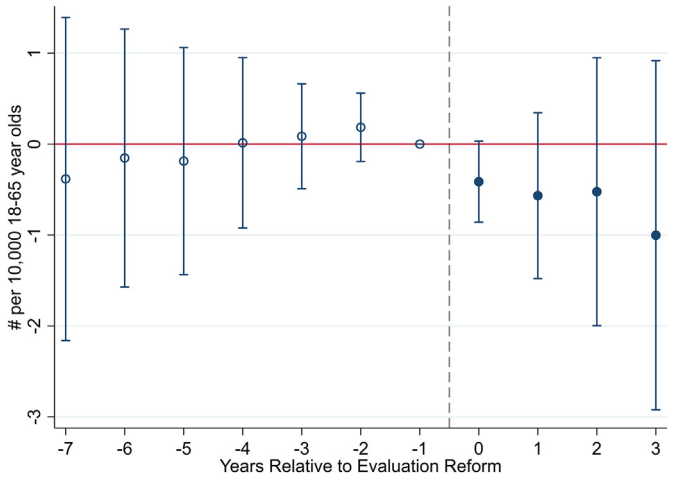

Fig. 8. Event study depicting effect of evaluation reforms on the number of graduates from graduate-level teacher preparation programs. Notes: Point estimates for years pre- (hollow dot) and post-reform (solid dot) and corresponding 95% confidence intervals are from models without controls reported in Appendix Table A2.

Table 8
The Effect of Education Policy Reforms on the Number of New Teaching Licenses.

	Separate Models	Joint Model
	(1)	(2)
Implement Evaluation	−2.400**	−2.599***
	(0.845)	(0.729)
Eliminate Tenure	−2.409	−0.510
	(1.800)	(1.693)
Increase Probationary Period	−1.452	−1.972
	(1.294)	(1.361)
Weaken Collective Bargaining	−0.560	−0.533
	(1.619)	(1.663)
Eliminate Mandatory Union Dues	−0.151	0.735
	(1.963)	(2.815)
Won Race to the Top Grant	0.075	0.485
	(1.268)	(1.153)
Number of Alt. Cert. Program Types	0.673	0.847*
	(0.468)	(0.366)
Implement Common Core Standards	−0.522	−0.108
	(1.585)	(1.229)
Basic Skills Licensure Tests	0.647	1.425
	(1.115)	(1.048)
Content Area Licensure Tests	−0.336	0.043
	(1.280)	(1.120)
Pedagogical Knowledge Licensure Tests	−1.048	−1.654
	(1.609)	(1.301)
Licensure Test Pass Rate	0.312**	0.384***
	(0.101)	(0.094)
Percent of Salary Withheld for Pension Fund	0.068	−0.125
	(0.428)	(0.401)
% change relative to state mean (Eval)	−15%	−16%
State Controls	Yes	Yes
n	750	750

Notes: +p<.10, *p<.05, **p<.01, ***P<.001. Cells report estimates and associated standard errors clustered at the state level in parentheses. See Table 2 notes for a complete list of time-varying state-level controls and Table 3 notes for a complete list of school-level controls. All models include state and region-by-year fixed effects. All models are weighted by the number 18-65 year old per 10,000 in a state. Estimates for licensure test pass rates and the percent of salary withheld for pension funds both correspond to a one percentage point change in the predictors. Sample sizes are rounded to the nearest tens place per IES disclosure guidelines. SOURCE: See Table 1.

models, we find direct and intuitive evidence that the number of new teacher licensures increases as the passing rates for licensure tests increase and the number of alternative certification programs increase. These results suggest states directly influence new teacher supply based on the minimum score they set for passing licensure tests and the number of alternative certification programs they approve.

9. Conclusion

Education policy over the past decade has focused considerable effort on improving human capital in schools through teacher accountability. These reforms, and the research base upon which they drew, were based on strong assumptions about how accountability would affect who decided to become a teacher. Counter to most assumptions, our findings document how teacher accountability reduced the supply of new teacher candidates by, in part, decreasing perceived job security, satisfaction and autonomy. This decline in supply had direct consequences for students, increasing the likelihood that schools could not fill vacant teaching positions. Even more concerning, effects on unfilled vacancies were concentrated in hard-to-staff schools that often serve larger populations of low-income students and students of color. At the same time, we find that evaluation reforms increased the quality of newly hired novice teachers by reducing the number of teachers that graduated from the least selective institutions. We find no evidence that evaluation reforms served to attract teachers who attended the most selective undergraduate institutions.

Together, these results provide new evidence of the competitiveness of the teacher labor market. The loss of non-pecuniary benefits due to accountability reforms was not offset by compensatory changes in teacher salaries at a national scale. Very few districts implemented merit pay programs of any significance despite efforts by the federal government and private philanthropists to more directly link teacher evaluation and compensation. Consistent with Rothstein's (2015) model-based predictions, the lack of meaningful offsetting increases in teacher base salaries or merit-based compensation caused a substantial fraction of would-be teachers to choose other professions or remain outside the labor market.

Enrollment in K-12 public schools in the United States is projected to increase by over a million students in the next decade, an increase of 2% from current levels. Even with new personalized learning technologies, it is hard to imagine a future where demand for classroom teachers is not increasing given expanding enrollments and a labor force where 29% of teachers are over the age of 50 (NCES, 2017). Understanding the consequences of education policy reforms on the supply and quality of new teachers will remain a key element of efforts to improve human capital in the education sector. This is particularly important for STEM subject areas with teacher shortages, as well as for the pipeline of new teachers needed to work in hard-to-staff schools in urban, rural, and low-income settings.

Declaration of competing interest

The authors declare no conflicts of interest.

Appendix A. Data sources for education policy reforms

A.1. Teacher evaluation reforms

We draw upon two systematic reviews of teacher evaluation reforms to create two binary measures that reflect the timing of evaluation reforms across states. We define "pass evaluation" as the year in which state legislatures pass a law reforming teacher evaluation systems and all following years. We define "implement evaluation" as the school year in which the state first implemented evaluation reforms state-wide and all following years. We code states as implementing evaluation even if some components of the new evaluation system were not yet in place or did not yet count towards teachers' formal evaluation scores.

Pass Year: We code passing evaluation reforms based on the calendar year the law was passed.

Implementation Year: We code implementing evaluation reforms based on the fall of the first academic year with state-wide implementation.
Sources: National Council on Teacher Quality (2016); Steinberg and Donaldson (2016). We prioritize the NCTQ report and use Steinberg and Donaldson (2016) to resolve ambiguities and address states with missing information.

A.2. Teacher tenure reforms

We compiled data on teacher tenure reforms from reviews of state statutes, case laws, new articles and prior literature. To capture the effect of the elimination of tenure on teacher labor supply, we created an indicator variable, *Eliminate Tenure,* which is equal to one when state eliminated tenure and all following years, and zero otherwise.
Note: Ohio and South Dakota passed laws eliminating tenure protections that were overturned the same year by state ballot referendums. Pennsylvania's tenure bill was vetoed by Governor. We do not code these three states as having eliminated tenure given that the legislation was overturned in the same calendar year. Georgia eliminated tenure in 2000 and later reinstated it in 2003. We do not code Georgia as eliminating tenure as their reform occurred prior to the start of our data panel. Advocates have filed lawsuits in California, Minnesota, and New York attempting to challenge the legality of teacher tenure although none of these suits were ultimately successful.
Year: We code tenure reforms based on the calendar year in which the state legislature passed the reform.
Sources:

- The National Council on Teacher Quality's (NCTQ) State Teacher Policy Yearbooks
- The Education Commission of the States' State Policy Database
- Students First: State Policy Report Cards

We resolved discrepancies between these sources using information from states' department of education websites, RTT federal grant applications, Lexis-Nexis searches of state and local newspapers, and conversations with academics and state Department of Education officials.

A.3. Probationary period increases

We collected data on probationary periods through a comprehensive review of state laws during our panel period. A total of 12 states extend their probationary period for tenure by at least one year during our panel period. We create a binary variable that takes on a value of one in the year a state increases the probationary period and all following years. This serves to eliminate the missing data problem caused by states that do not have probationary periods or teacher tenure.
Note: Hawaii decreased the probationary period in 2008 from two to one years and then raised to two again in 2009. It then raised the probationary period to three in 2013. We only code this later raise in our data.
Year: We code these changes as occurring in the year in which the legislation was passed. It appears these laws all went into effect immediately upon being enacted.
Source: Authors' own research based on news articles and state legislative documents.

A.4. Collective bargaining reforms

We compiled data on collective bargaining reforms through a comprehensive review of state laws during our panel period. A total of six states passed legislation that eliminated mandatory collective bargaining with public teacher unions or substantially restricted the scope of which aspects of teachers' contracts were subject to collective bargaining (ID [later reversed], IN, MO, NM, TN, and WI). Ohio also made collective bargaining illegal in 2011, but the law was overturned that same year in a statewide referendum. We control for a time-varying indicator of whether collective bargaining is not mandatory in a given state.
Year: We code collective bargaining reforms based on the calendar year in which the state legislature passed the reform.
Source: Authors' own research based on news articles and state legislative documents.

A.5. Right to work reforms (eliminate mandatory union dues)

We collected data on probationary periods through a comprehensive review of state laws during our panel period. A total of four states ended mandatory union dues during our panel period. These laws were challenged in court in three of these states (IN, WI, and WV), but were ultimately upheld.
Year: We code right to work laws as occurring in the year in which the legislation was passed. It appears these laws all went into effect immediately upon being enacted.
Source: Authors' own research based on news articles and state legislative documents.

A.6. Race to the top winners

A full list of the timing and dates of race to the top winners is available from the U.S. Department of Education. We code this time-varying variable as taking on a value of one in the year grant awards were announced and all following years.
Year: We code Race to the Top winners based on the calendar year in which the award was announced by the U.S. Department of Education.
Source: U.S. Department of Education.
https://www2.ed.gov/programs/racetothetop/awards.html

A.7. Alternative certification programs

We measure the availability of alternative certification programs using data maintained by the U.S. Department of Education Office of Title II. These data track characteristics of alternative certification programs in operation across each state including the year in which each program was first implemented. We use these data to reconstruct a panel dataset from 2002 to 2016 with counts of the number of alternative certification programs operating in each state in each year.
Year: We code alternative certification programs based on the implementation year (calendar vs. academic unspecified) provided by U.S. Department of Education, Higher Education Act Title II State Report Card System.
Source: U.S. Department of Education Office of Title II.
https://title2.ed.gov/Public/DataTools/Tables.aspx

A.8. Common core state standards

We compiled data on the implementation of the Common Core State Standards (CCSS) from the Common Core State Website. We compiled data on the timing and duration of a state's adoption of the Common Core State Standards (CCSS) from several news sources listed below. We constructed a time-varying indicator for the school years in which states implemented the common core state standards statewide. Oklahoma and Indiana had initially adopted the standards, but dropped them before implementing them statewide. South Carolina implemented the CCSS for the 2014–15 year, but then abandoned the standards the following year. Some states such as Pennsylvania have renamed their standards and modified them. States that have maintain at least 85% of the CCSS maintain their CCSS status.

Year: We code CCSS implementation based on the fall school year.

Source: Common Core State Website.

http://www.corestandards.org/standards-in-your-state/

We use a range of news article to update these data for states that dropped CCSS.

A.9. New teacher licensure tests

We construct time-varying indicators for different types of licensure exams that states can require based on tables from the annual Digest of Education Statistics compiled by the National Center for Education Statistics (NCES). Each year the digest contains a table titled "States requiring testing for initial certification of teachers, by skills or knowledge and state." We include indicators for three types of licensure exams coded: basic skills exams, subject exams, and professional knowledge exams. The basic skills exam is a test of fundamental reading, writing and mathematics skills. The subject exams test content knowledge in the area for which a teacher will receive licensure. The professional knowledge exam tests knowledge of pedagogy in areas such as educational psychology, classroom management, lesson planning and evaluation, and assessment. When data was listed as missing in the tables we coded it as zero for not having adopted a new licensure test.

Note: Some states, prior to instituting a uniform licensure test, used other exams such as an "Institution's Exit Exam" or "Subject matter exam or completion of an approved subject matter program" as proxies. We do not code these licensure tests.

Year: We code licensure data based on the fall school year a new law was adopted.

Source: Annual Digest of Education Statistics compiled by the National Center for Education Statistics (NCES).

https://nces.ed.gov/programs/digest/d11/tables/dt11_179.asp

The source data for these tables come from National Association of State Directors of Teacher Education and Certification (NASDTEC) Manuel on the Preparation and Certification of Educational Personnel in the United States and Canada States.

A.10. Teacher licensure exam passing rates

We measure teacher passing rates for each state for each year using the number of licensure exams taken and passers in a given state in a given year, and then calculating the share of tests that scored at or above the passing threshold. This measure is an average across all licensure tests required for an initial teaching credential in a state, which vary by state. Approximately 74% of these tests were administered by ETS, 20% were administered by Pearson, and the rest were administered by individual states. Because each state sets its own cut score for passing each exam, the cut scores vary across states and across exams. Note that some states, particularly in the earlier years, have missing data due to differences in assessment requirements and systems and processes for state reporting. We impute the state average pass rate across the panel for missing years and include an indicator for missing data.

Year: The report year provided in the Title II database reflects the year that follows the spring academic year. We recode this to the prior calendar year, the year in which most teachers were likely taking the test (January – July) to start teaching in the prior academic year.

Source: The data on state by-year licensure exam pass rates come from the Title II office, who obtained data from the states. States and teacher preparation providers work directly with testing companies to submit lists of enrolled students and program completers. The testing companies match those lists to produce the pass rates for each assessment and provider, and there are several rounds of data verification. Once the data are verified by all parties, the testing companies package the data and send them to the Department of Education to upload into the Title II reports.

https://title2.ed.gov/Public/Report/DataFiles/DataFiles.aspx?p=5_01

Data starting in 2010–2011 are available online. We obtained earlier data in aggregate form from Westat which provides technical assistance and manages the Title II data.

A.11. Teacher pensions

We measure teacher pension contributions as the required employee contribution rate, which captures the percent of total wages that teachers must contribute towards pension funds. We control for a continuous measure of the percentage of teachers' salary that they contribute towards pension funds in each state and year.

Year: We use the year (likely calendar) provided in the Public Plans Database.

Source: Annual data on employee contribution rates for state and locally sponsored teacher pension funds comes from the Public Plans Data maintained by the Center for Retirement Research at Boston College.

https://crr.bc.edu/data/public-plans-database/

A.12. Certified public accountants

Our state-by-year cohort counts are collected by the National Association of State Boards of Accountancy (NASBA) in partnership with The American Institute of Certified Public Accountants (AICPA). All candidates who begin the CPA examination process (begin any one of the four sub-exams) in the same calendar year are members of a state-year cohort. Requirements for CPA candidates are similar to those for teaching candidates – while all CPA candidates are required to have *at least* a bachelor's degree and complete a set number of accounting courses, the CPA examination/licensure requirements vary from state-to-state.

Table A1
Education Policy Reforms by State, 2002-2016.

	Pass Evaluation	Implement Evaluation	Eliminate Tenure	Increase Probationary Period	Weaken Collective Bargaining	Eliminate Mandatory Union Dues	Won Race to the Top Grant	Implement Common Core Standards	Basic Skills Licensure Tests	Content Area Licensure Tests	Pedagogical Knowledge Licensure Tests
Alabama	2013	2016						2013	2003	2005	2005
Alaska	2012	2016							2006	2016	
Arizona	2010	2013					2011	2013			
Arkansas	2011	2014						2013	2002		
California								2014			
Colorado	2010	2013					2011	2013		2007	
Connecticut	2012	2014						2013			2005
Delaware	2010	2012					2010	2012		2005	
Florida	2011	2011	2011				2010				
Georgia	2012	2014					2010	2014			
Hawaii	2010	2013		2013			2010	2013			
Idaho	2011	2014	2011		2011			2013		2005	2005
Illinois	2010	2016					2011	2013	2005	2005	2003
Indiana	2011	2014			2005	2012					
Iowa								2012	2003	2015	2015
Kansas	2012	2014	2014					2013		2004	
Kentucky	2013	2014					2011	2011	2008	2003	2003
Louisiana	2010	2012	2012	2012			2011	2013			
Maine	2012	2016		2011				2012	2015	2015	2015
Maryland	2010	2013		2010			2010	2013			
Massachusetts	2011	2013					2010	2013	2005	2005	
Michigan	2010	2011		2011	2011	2013		2012			
Minnesota	2011	2014								2005	2002
Mississippi	2012	2014						2013	2015	2015	2015
Missouri	2011	2013			2005			2014	2005		2015
Montana								2013		2016	
Nebraska	2012								2006	2016	
Nevada	2011	2015		2011				2013	2015	2006	
New Hampshire	2013	2016		2011				2014			
New Jersey	2012	2013		2012			2011	2013	2015	2015	
New Mexico	2011	2013						2013		2003	
New York	2010	2012		2015			2010	2013	2015	2005	
North Carolina	2010	2011	2013				2010	2012	2015	2015	2015
North Dakota	2011	2015						2013	2015	2008	2015
Ohio	2011	2013		2009			2010	2013		2006	2006
Oklahoma	2010	2013									2004
Oregon	2011	2013						2014	2006	2006	
Pennsylvania	2012	2013					2011	2013			
Rhode Island	2010	2012					2010	2013	2015	2015	2005
South Carolina	2013	2014									
South Dakota	2010	2015						2014	2007		2007
Tennessee	2010	2011		2011	2011		2010	2013	2005		
Texas	2013	2016							2012	2012	2012
Utah	2012	2015						2013	2016	2005	
Vermont								2013	2002	2016	
Virginia	2011	2012									2005
Washington	2012	2015		2010				2014	2005	2005	
West Virginia	2011	2013				2016		2014			
Wisconsin	2012	2014			2011	2011		2014		2005	
Wyoming	2012							2014			

Notes: See Appendix A for details about each measure. Our time-varying measures account for reforms that were later reversed. States often passed several laws over multiple year related to evalution reforms. Our pass evaluation code represent the laws we judge to be most consequential, not necessarily the first. Georgia eliminated tenure in 2000 and reinstituted it in 2003. Idaho revoked tenure laws in 2011 but reinstated them in 2012. In 2012, Louisiana made tenure exceedingly difficult to receive by requiring teachers to have a highly-effective rating five out of six consecutive years. This table does not contain data on two other non-binary education reforms for which we control, the number of alternative certification program types and the percent of salary withheld for pension funds.

Table A2
The Effect of Teacher Evaluation Reforms from Event Study Models.

	Number of Licenses		At least one unfilled vacancy in a school		Undergraduate Institution with Competitive Admissions		The Number of Program Completers from Graduate Teacher Preparation Programs		Agree Worry about Job Security	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
$r=-7$ or <	0.516	1.736	−0.017	−0.022+	−0.049	0.039	−0.383	0.268	−0.202*	−0.179*
	(1.755)	(1.670)	(0.015)	(0.012)	(0.042)	(0.044)	(0.907)	(0.755)	(0.079)	(0.086)
$r=-6$	0.056	0.768	−0.009	−0.013	−0.135**	−0.073	−0.153	0.326	−0.152+	−0.091
	(1.345)	(1.369)	(0.018)	(0.015)	(0.042)	(0.047)	(0.724)	(0.592)	(0.089)	(0.098)
$r=-5$	0.540	1.347	−0.005	−0.002	0.036	0.082*	−0.187	0.233	−0.222**	−0.195*
	(1.221)	(1.180)	(0.012)	(0.008)	(0.033)	(0.039)	(0.637)	(0.576)	(0.075)	(0.076)
$r=-4$	0.955	1.696	−0.036*	−0.044**	−0.093+	−0.003	0.014	0.233	−0.209*	−0.197*
	(1.263)	(1.186)	(0.018)	(0.015)	(0.050)	(0.055)	(0.478)	(0.486)	(0.087)	(0.096)
$r=-3$	0.845	1.481	−0.000	−0.013	−0.051	0.056	0.086	0.184	−0.215***	−0.213*
	(1.104)	(1.060)	(0.013)	(0.012)	(0.046)	(0.046)	(0.294)	(0.275)	(0.059)	(0.080)
$r=-2$	2.281+	2.725*	0.005	0.009	−0.097*	−0.075+	0.185	0.241	−0.178*	−0.133+
	(1.137)	(1.135)	(0.011)	(0.011)	(0.042)	(0.042)	(0.192)	(0.163)	(0.070)	(0.073)
$r=0$	−0.802	−1.040	0.008	−0.001	−0.018	0.065	−0.413+	−0.511*	0.003	0.023
	(0.704)	(0.728)	(0.014)	(0.012)	(0.037)	(0.041)	(0.228)	(0.191)	(0.053)	(0.063)
$r=1$	−2.679*	−3.189**	0.029*	0.013	0.086	0.184**	−0.567	−0.736+	0.004	−0.025
	(1.071)	(1.055)	(0.013)	(0.013)	(0.057)	(0.056)	(0.465)	(0.389)	(0.071)	(0.091)
$r=2$	−3.558*	−4.193**	0.051**	0.041**	−0.051	−0.007	−0.524	−0.766	0.089+	0.097
	(1.371)	(1.292)	(0.017)	(0.015)	(0.044)	(0.048)	(0.752)	(0.604)	(0.051)	(0.073)
$r=3$ or >	−4.828*	−6.043**	0.043**	0.029*	0.142*	0.163**	−1.003	−1.370+	0.166*	0.103
	(2.096)	(1.872)	(0.013)	(0.013)	(0.063)	(0.055)	(0.980)	(0.804)	(0.073)	(0.089)
School Controls	No	No	No	Yes	No	Yes	No	No	No	Yes
State Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Policy Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
n	750	750	28,610	28,610	5,800	5,800	750	750	5,800	5,800

Notes: +p<.10, *p<.05, **p<.01, ***P<.001. Cells report estimates and associated standard errors clustered at the state level in parentheses. See Table 2 notes for a complete list of time-varying state and policy controls and Table 3 notes for a complete list of school controls. All models include state and region-by-year fixed effects. Sample sizes are rounded to the nearest tens place per IES disclosure guidelines. SOURCE: See Table 1.

Table A3
The Effect of Teacher Evaluation Reforms on the Competitiveness of New Teachers' Undergraduate Institutions.

	Less Competitive or higher	Competitive or higher	Very Competitive or higher	Highly Competitive or higher	Most Competitive or higher
	(2)	(3)	(4)	(5)	(6)
Implement Evaluation	0.020	0.081**	0.031	0.004	−0.007
	(0.015)	(0.028)	(0.062)	(0.036)	(0.016)
n	5,800	5,800	5,800	5,800	5,800
School Controls	Yes	Yes	Yes	Yes	Yes
State Controls	Yes	Yes	Yes	Yes	Yes
Policy Controls	Yes	Yes	Yes	Yes	Yes

Notes: +p<.10, *p<.05, **p<.01, ***P<.001. Cells report estimates and associated standard errors clustered at the state level in parentheses. See Table 2 notes for a complete list of time-varying state and policy controls and Table 3 notes for a complete list of school controls. All models include state and region-by-year fixed effects. Sample sizes are rounded to the nearest tens place per IES disclosure guidelines. SOURCE: See Table 1.

Table A4
The Effect of Teacher Evaluation Reforms on New Teacher Working Conditions.

	(1)	(2)	(3)	(4)
	Panel A. Worry About Job Security			
Implement Evaluation	1.892***	2.038***	1.572*	1.791**
	(0.320)	(0.354)	(0.337)	(0.405)
Implement Evaluation * Trend			1.066	1.034
			(0.073)	(0.070)
Trend			1.033	1.035
			(0.022)	(0.028)
	Panel B. Job Satisfaction			
Implement Evaluation	0.579***	0.523***	0.636*	0.685+
	(0.073)	(0.083)	(0.114)	(0.134)
Implement Evaluation * Trend			0.968	0.866*
			(0.055)	(0.061)
Trend			0.983	0.983
			(0.021)	(0.023)
	Panel C. Teacher Cooperation			
Implement Evaluation	0.673**	0.582***	0.475***	0.455***
	(0.088)	(0.077)	(0.090)	(0.077)

Table A4 (*continued*)

	(1)	(2)	(3)	(4)
Implement Evaluation * Trend			1.181*	1.107
			(0.078)	(0.075)
Trend			1.022	1.041+
			(0.020)	(0.022)
	Panel D. Control over Selecting Textbooks and Instructional Materials			
Implement Evaluation	0.997	0.958	1.141	1.001
	(0.165)	(0.137)	(0.250)	(0.216)
Implement Evaluation * Trend			0.960	0.991
			(0.060)	(0.061)
Trend			0.971	0.987
			(0.032)	(0.027)
	Panel E. Control over Content, Topics, Skills Taught			
Implement Evaluation	0.765+	0.736+	0.694*	0.590**
	(0.118)	(0.115)	(0.111)	(0.099)
Implement Evaluation * Trend			1.036	1.072
			(0.051)	(0.056)
Trend			1.016	1.050
			(0.043)	(0.035)
	Panel F. Control over Teaching Techniques			
Implement Evaluation	0.772	0.656*	0.669**	0.569**
	(0.124)	(0.110)	(0.096)	(0.106)
Implement Evaluation * Trend			1.048	0.998
			(0.061)	(0.051)
Trend			1.031	1.071*
			(0.029)	(0.033)
n	6,460	6,460	6,460	6,460
School Controls	No	Yes	No	Yes
State Controls	No	Yes	No	Yes
Policy Controls	No	Yes	No	Yes

Notes: +p<.10, *p<.05, **p<.01, ***P<.001. Cells report estimates and associated standard errors clustered at the state level in parentheses. See Table 2 notes for a complete list of time-varying state and policy controls and Table 3 notes for a complete list of school controls. All models include state and region-by-year fixed effects. Sample sizes are rounded to the nearest tens place per IES disclosure guidelines. SOURCE: See Table 1.

Table A5
Alternative Model Specifications for the Effect of Teacher Evaluations on the Number of New Teaching Licenses.

Alternative Specification	Year state legislatures passes evaluation reforms		Replacing region-by-year with year fixed effects		Restrict sample to only 29 states that only passed teacher evaluation reforms		Include endogenous controls for teacher demand		Pre a post trends for concurrent accountability reforms and policy controls	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Evaluation	−1.898**	−3.021***	−2.421**	−1.725+	−4.169***	−2.504*	−2.733***	−1.276	−1.680*	0.222
	(0.652)	(0.711)	(0.820)	(0.977)	(1.111)	(1.074)	(0.692)	(0.888)	(0.738)	(0.751)
Evaluation * Trend		−2.239***		−0.943+		−1.383*		−1.135*		−2.192***
		(0.422)		(0.523)		(0.623)		(0.453)		(0.394)
Trend		1.023***		0.368**		0.075		0.001		0.482*
		(0.205)		(0.136)		(0.298)		(0.212)		(0.232)
% change relative to state mean	−12%		−15%		−26%		−17%		−10%	
State Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Policy Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
n	750	750	750	750	410	410	750	750	750	750

Notes: +p<.10, *p<.05, **p<.01, ***P<.001. Cells report estimates and associated standard errors clustered at the state level in parentheses. See Table 2 notes for a complete list of time-varying state and policy controls. Endogenous controls for teacher labor demand include the number of novice public school teachers, logged public and private school enrollment, and public school pupil-teacher ratio. We impute data for years in which the number of novice public school teachers and private school enrollment are not available using linear interpolation. Sample sizes are rounded to the nearest tens place per IES disclosure guidelines. SOURCE: See Table 1.

Table A6
Alternative Model Specifications for the Effect of Teacher Evaluations on the Probability Schools Have at least One Unfilled Teaching Position.

Alternative Specification	Year state legislatures passes evaluation reforms		Replacing region-by-year with year fixed effects		Restrict sample to only 29 state that only passed teacher evaluation reforms		Include endogenous controls for teacher demand		Pre a post trends for concurrent accountability reforms and policy controls	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Evaluation	0.004	−0.004	0.020**	0.011	0.053***	−0.006	0.025**	0.009	0.032***	0.016
	(0.006)	(0.006)	(0.007)	(0.009)	(0.012)	(0.026)	(0.008)	(0.008)	(0.009)	(0.010)
Evaluation * Trend		0.013***		0.004		0.017+		0.006+		0.006
		(0.003)		(0.003)		(0.009)		(0.003)		(0.005)
Trend		0.001		0.001		0.006***		0.003**		0.004*
		(0.002)		(0.001)		(0.001)		(0.001)		(0.002)
School Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Policy Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
n	28,610	28,610	28,610	28,610	16,880	16,880	28,610	28,610	28,610	28,610

Notes: +p<.10, *p<.05, **p<.01, ***P<.001. Cells report estimates and associated standard errors clustered at the state level in parentheses. See Table 2 notes for a complete list of time-varying state and policy controls and Table 3 notes for a complete list of school controls. All models include state and region-by-year fixed effects and use appropriate sampling weights provided by the SASS/NTPS. Endogenous controls for teacher labor demand include the number of novice public school teachers, logged public and private school enrollment, and public school pupil-teacher ratio. We impute data for years in which the number of novice public school teachers and private school enrollment are not available using linear interpolation. Sample sizes are rounded to the nearest tens place per IES disclosure guidelines. SOURCE: See Table 1.

Table A7
Alternative Weighting Approaches for Modeling the Effect of Teacher Evaluations on the Number of New Teaching Licenses.

Outcome specification	Scaled per number of 18-65 year olds per 10,000		Scaled per number of 22-25 year olds per 1,000		Scaled per number of 18-65 year olds per 10,000		Logged, controlling for logged number of 18-65 year olds per 10,000	
Weights	Number of 18-65 year olds per 10,000		Number of 22-25 year olds per 1,000		No weights		No weights	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Implement Evaluation	−2.996***	−1.256	−3.654***	−1.576	−2.880**	−1.729+	−0.202**	−0.097
	(0.722)	(0.907)	(0.861)	(1.044)	(0.927)	(0.898)	(0.061)	(0.065)
Implement Evaluation * Trend		−1.468**		−1.789**		−1.225**		−0.097**
		(0.456)		(0.527)		(0.404)		(0.034)
Trend		0.188		0.249		0.214		0.010
		(0.177)		(0.207)		(0.218)		(0.014)
% change relative to state mean	−18%		−22%		−18%			
State Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Policy Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
n	750	750	750	750	750	750	750	750

Notes: +p<.10, *p<.05, **p<.01, ***P<.001. Cells report estimates and associated standard errors clustered at the state level in parentheses. See Table 2 notes for a complete list of time-varying state and policy controls. Sample sizes are rounded to the nearest tens place per IES disclosure guidelines. SOURCE: See Table 1.

Table A8
Alternative Model Specifications for the Effect of Teacher Evaluations on the Number of Graduates from M.A. Granting Teacher Preparation Programs.

Alternative Specification	Year state legislatures passes evaluation reforms		Replacing region-by-year with year fixed effects		Restrict sample to only 29 state that only passed teacher evaluation reforms		Include endogenous controls for teacher demand		Pre a post trends for concurrent accountability reforms and policy controls	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Evaluation	0.341	0.335	−0.501	−0.279	−0.527	−0.680	−0.680*	−0.400+	−0.187	−0.384
	(0.474)	(0.517)	(0.329)	(0.264)	(0.501)	(0.519)	(0.311)	(0.227)	(0.332)	(0.281)
Evaluation * Trend		−0.482*		−0.195		−0.143		−0.164		0.073
		(0.232)		(0.181)		(0.456)		(0.173)		(0.261)
Trend		0.110		0.038		0.170		−0.050		0.141+
		(0.109)		(0.075)		(0.144)		(0.085)		(0.084)
% change relative to state mean	5%		−8%		−8%		−11%		−3%	
State Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Policy Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
n	750	750	750	750	420	420	750	750	750	750

Notes: +p<.10, *p<.05, **p<.01, ***P<.001. Cells report estimates and associated standard errors clustered at the state level in parentheses. See Table 2 notes for a complete list of time-varying state and policy controls. Endogenous controls for teacher labor demand include the number of novice public school teachers, logged public and private school enrollment, and public school pupil-teacher ratio. We impute data for years in which the number of novice public school teachers and private school enrollment are not available using linear interpolation. Sample sizes are rounded to the nearest tens place per IES disclosure guidelines. SOURCE: See Table 1.

Table A9
Alternative Weighting Approaches for Modeling the Effect of Teacher Evaluations on the Number of Graduates from M.A. Granting Teacher Preparation Programs.

Outcome specification	Scaled per number of 18-65 year olds per 10,000		Scaled per number of 22-25 year olds per 1,000		Scaled per number of 18-65 year olds per 10,000		Logged, controlling for logged number of 18-65 year olds per 10,000	
Weights	Number of 18-65 year olds per 10,000		Number of 22-25 year olds per 1,000		No weights		No weights	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Implement Evaluation	−0.674*	−0.324	−0.918*	−0.453+	−0.516+	−0.413+	−0.073+	−0.019
	(0.322)	(0.215)	(0.374)	(0.251)	(0.280)	(0.224)	(0.039)	(0.032)
Implement Evaluation * Trend		−0.287		−0.389+		−0.191		−0.041
		(0.183)		(0.204)		(0.191)		(0.027)
Trend		0.031		0.047		0.069		−0.000
		(0.083)		(0.095)		(0.087)		(0.014)
% change relative to state mean	−10%		−14%		−8%			
State Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Policy Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
n	750	750	750	750	750	750	750	750

Notes: +p<.10, *p<.05, **p<.01, ***P<.001. Cells report estimates and associated standard errors clustered at the state level in parentheses. See Table 2 notes for a complete list of time-varying state and policy controls. Sample sizes are rounded to the nearest tens place per IES disclosure guidelines. SOURCE: See Table 1.

Table A10
The Effect of Teacher Evaluation Reforms on Public Teacher Salaries.

	NCES Average Salary		ACS Adjusted Average Salary	
	(1)	(2)	(3)	(4)
Implement Evaluation	−330.652	−173.864	−509.883+	−152.010
	(499.984)	(388.157)	(275.208)	(261.144)
Implement Evaluation * Trend		237.096		−195.319
		(245.300)		(191.060)
Trend		−334.939*		−127.370
		(126.930)		(107.655)
% change relative to state mean	−1%		−1%	
State Controls	Yes	Yes	Yes	Yes
Policy Controls	Yes	Yes	Yes	Yes
n	750	750	600	600

Notes: +p<.10, *p<.05, **p<.01, ***P<.001. Cells report estimates and associated standard errors clustered at the state level in parentheses. See Table 2 notes for a complete list of time-varying state and policy controls. All models include state and region-by-year fixed effects and are weighted by the total number of Full Time Equivalent (FTE) teachers in each state and year. Sample sizes are rounded to the nearest tens place per IES disclosure guidelines. SOURCE: See Table 1.

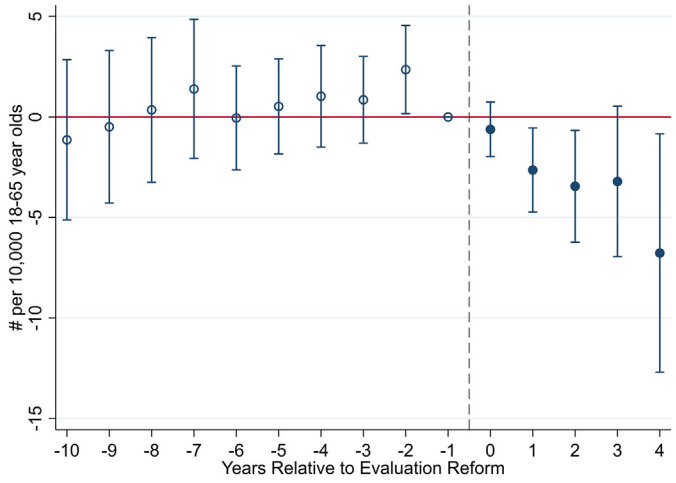

Fig. A1. Event Study Depicting Effect of Evaluation Reforms on the Number of New Teaching Licenses with Alternative Minimum and Maximum Years. Notes: Point estimates for years pre- (hollow dot) and post-reform (solid dot) and corresponding 95% confidence intervals are from models without controls reported in Appendix Table A2.

Appendix B. *B.1.* Roy model of occupational choice

To illustrate the potential effects of teacher evaluation reforms on the supply and quality of prospective teachers, we adapt a simple Roy (1951) model of occupational choice in which accountability reforms increase the relative costs associated with teaching. Individuals choose between a career teaching in public schools (hereafter teaching), T, or an alternative occupation, A, that represents all outside labor market options, by choosing the occupation that maximizes expected earnings, w. Earnings in teaching and the alternative occupation are given by:

$$w_T = \mu_T + \eta_T \nu - Cg(\nu)$$

$$w_A = \mu_A + \nu$$

where μ_T and μ_A denote average earnings in the teaching and alternative occupation respectively, ν denotes the individual's ability, which is continuously distributed with mean zero and variance σ^2_ν, and η_T is the return to ability in the teaching sector relative to the alternative occupation. Assuming $0 < \eta_T < 1$, earnings will be more compressed in the teaching sector.[17]

The term C denotes costs common to all teachers. These costs may include the opportunity costs of earning a teaching certificate from an educator preparation program and passing state teacher licensure exams. We conceptualize evaluation reforms as increasing C through the monetized costs of declines in job security and/or job satisfaction in teaching relative to the alternative occupation. The term $g(\nu)$ allows costs to depend on ability. We assume $g(\nu)$ is from the family of exponential functions with: $g(\nu) > 0$, $g_\nu < 0$, $g_{\nu\nu} > 0$, $\lim_{\nu \to -\infty} g(\nu) = \infty$, $\lim_{\nu \to \infty} g(\nu) = 0$.[18] The condition $g_\nu < 0$ implies the relative costs associated with teaching decline with ability.[19]

An individual chooses teaching as a career if $w_T > w_A$, implying that the individual indifferent between a career in teaching and the alternative occupation is characterized by the implicit function:[20]

$$F(\nu, C, \eta_T, \mu_T, \mu_A) = \mu_T - \mu_A - Cg(\nu) - \nu(1 - \eta_T) = 0. \tag{1}$$

Fig.B1 illustrates the effect of an increase in the relative costs of teaching. Note that the concavity of $F(\nu, C, \eta_T, \mu_T, \mu_A)$ with respect to ν implies there is both a high- *and* low-ability individual on the margin between teaching and the alternative occupation. Denoting the marginal low- and high-ability individuals as ν_L *and* ν_H respectively, the change in $\nu_j \in (\nu_L, \nu_H)$ due to a change in the relative costs associated with teaching is:

$$\frac{\partial \nu_j}{\partial C} = \frac{g(\nu_j)}{-Cg_{\nu_j} - (1 - \eta_T)}. \tag{2}$$

Because the numerator of (2) is strictly positive, the sign of (2) depends on the sign of the denominator, which is the slope of the implicit function given by (1) evaluated at the roots, ν_j. Consequently, it follows that, $\frac{\partial \nu_L}{\partial C} > 0$ and $\frac{\partial \nu_H}{\partial C} < 0$, causing ν_L to shift right to ν_L' and ν_H to shift left to ν_H'. As a result, both the share of high- and low-ability individuals that choose teaching as a career declines, leading to an unambiguous reduction in the supply of prospective teachers but an ambiguous change in teacher quality.[21] Note that our model also predicts that ν_L shifts to the right by more than ν_H shift to the left. Intuitively, because costs decline with ability, costs rise more for the marginal low-ability individual than for the marginal high-ability individual. Nevertheless, the net effect of these changes on teacher quality remains ambiguous since it depends on the density of individuals close to the marginal low- and high-ability teacher.

B.2. Changes in the return to ability

While teacher evaluation reforms may increase the perceived costs associated with teaching, merit pay schemes based on evaluation ratings could attract more qualified teachers into the profession. In the context of our model, merit pay can be thought of as an increase η_T or an increase in the returns to ability in the teaching sector. Solving for the change in the marginal ability individual due to a change in η_T yields:

$$\frac{\partial \nu_j}{\partial \eta_t} = \frac{-\nu_j}{-Cg_{\nu_j} - (1 - \eta_T)}. \tag{3}$$

To sign (3) note that the denominator is the same as the denominator of (2), implying it is positive for ν_L and negative for ν_H. The sign of (3) therefore depends on the numerator, which solely depends on ability evaluated at ν_L and ν_H. Empirical evidence on the ranking of teachers within the ability distribution of college graduates suggests that on average teachers rank around the 40th to 50th percentile relative to their peers, implying that ν_L is likely negative and ν_H positive when ν is standardized with mean zero.[22] An increase in η_T therefore causes the share of low-ability individuals that

[17] For simplicity we assume that the type of ability that is valued in the teaching and non-teaching sectors is perfectly positively correlated. More realistically, the types of skills valued in the two sectors likely differs to some extent implying that ν may differ across sectors. Nevertheless, if ability is valued in both sectors and the correlation between the type of ability that is valued in both sectors is sufficiently strong, then relaxing the assumption that ν is the same across sectors leads to predictions that are qualitatively the same as those based on our simplifying assumption.

[18] Examples of functional forms that satisfy these assumptions are $g(\nu) = e^{-\rho\nu}$ and $g(\rho, \nu) = \rho^\nu$, where the parameter $0 < \rho < 1$ determines the degree to which costs differ by ability.

[19] A special case of our model occurs when the relative costs associated with teaching are the same for all individuals, implying $Cg(\nu) = C$. In that case it is trivial to show that an increase in C, leads to a reduction in the supply of prospective teachers and a decline in the average ability of teachers. Intuitively, when the relative costs are the same for everyone, individuals on the margin between teaching and the alternative occupation are the highest ability prospective teachers. Consequently, when costs increase, the supply and average ability of individuals choosing a career in teaching declines.

[20] Note that $\lim_{\nu \to -\infty} F(\nu, C, \eta_T, \mu_T, \mu_A) = -\infty$ and $\lim_{\nu \to \infty} F(\nu, C, \eta_T, \mu_T, \mu_A) = -\infty$.

[21] More formally, the share of individuals that choose a career in teaching is given by $\Pr(T) = \int_{\nu_L}^{\nu_H} f(\nu) d\nu$ and the effect of an increase in the relative costs associated with teaching on the supply of prospective teachers is: $\frac{\partial \Pr(T)}{\partial C} = \frac{\partial \nu_H}{\partial C} f(\nu_H) - \frac{\partial \nu_L}{\partial C} f(\nu_L) < 0$.

[22] For example, Master et al. (2018) find that in 2000 and 2008, college graduates entering the teaching profession ranked around the 42nd and 48th percentiles, respectively, on SAT and ACT scores relative to their peers. For more evidence see Corcoran et al. (2004), Goldhaber and Walch (2014), and Lankford et al. (2014).

choose to teach to decline, $\frac{\partial v_L}{\partial \eta_t} > 0$, and the share of high-ability individuals that choose to teach to increase, $\frac{\partial v_H}{\partial \eta_t} > 0$. Consequently, the introduction of merit pay causes average teacher quality to increase while leading to an ambiguous change in the supply of prospective teachers.

In summary, when the relative costs associated with teaching increase due to the implementation of teacher evaluation reforms, both the share of high- and low-ability individuals that choose teaching as a career declines, leading to an overall reduction in the supply of prospective teachers and an ambiguous change in teacher quality. However, if evaluation reforms simultaneously increase both the relative cost of teaching and the degree to which earnings depend on ability (i.e., increase η_T), the share of high-ability individuals that choose to teach may increase. As a result, the quality of prospective teachers may increase but the effect on the share of individuals that choose to teach is ambiguous.

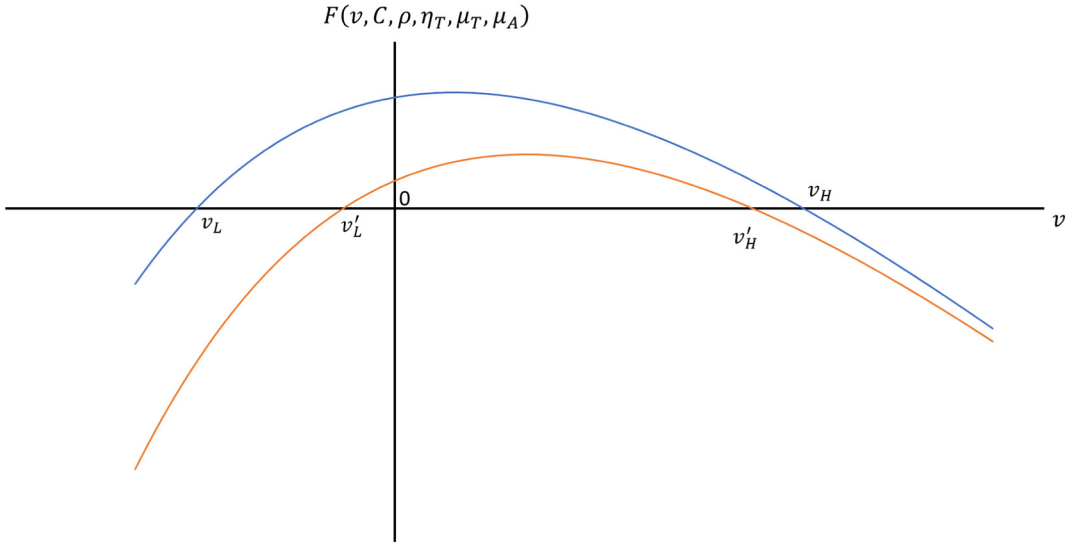

Fig. B1. An Increase in the Relative Costs of Teaching when Costs Vary by Ability.

Appendix C. Identifying graduates of teacher preparation programs and by subject areas using CIPS codes

The Integrated Postsecondary Education Data System (IPEDS) includes the results of eleven surveys conducted annually by the U.S. Department of Education's National Center for Education Statistics (NCES). All institutions must submit race and ethnicity data if they receive, are applicants for, or expect to be applicants for federal financial assistance as defined under the DOE's regulations implementing Title VI of the Civil Rights Act of 1964, Title IX of the Education Amendments of 1972, and the Carl D. Perkins Education Act. These surveys thus capture the substantive universe of postsecondary schools, colleges, and universities.

In this paper, we use the entire "Completions" survey available for download on the IPEDS website. Each observation in the file corresponds to the completers of a particular academic program, identified by classification of instruction program (CIP) code, and the award level (e.g., bachelor's degree, master's degree, etc.) of that particular program for each reporting institution.

The NCES has devised a six-digit taxonomy for organizing academic programs called CIP codes. The first two digits of the code correspond to a broad area of study. For instance, all majors under the two-digit CIP category "13" are majors within the education field. These majors include not only education programs designed to prepare individuals to be teachers, but also teaching assistants preparation programs (13.1501) and programs for education program evaluators (e.g., Education Evaluation and Research, 13.0601), among others. The next two digits in the CIP code cluster majors with similar instructional content within the board area of study. All majors with a 13.13 CIP code sequence, for example, are "teacher education or professional development majors within specific subject areas" within the broad field of education. The final two digits are unique to each specific major that falls within the specific subject area. For example, a "structural engineering" major has a 14.0803 CIP code ("14" corresponds to engineering, "08" correspond to civil engineering, and the final "03" is unique to structural engineering). NCES adds and removes CIP codes regularly. We created consistent categories for all six-digit codes across the panel using crosswalks supplied by NCES.[23]

Based on conversations with IPEDS-reporting institutions and the IPEDS Help Desk Staff, we identify teacher preparation programs based on the following CIP codes[24]:

- Education, General: 13.0101
- Bilingual, Multilingual, and Multicultural Education 13.0201–12.0299
- Curriculum and Instruction: 13.0301
- Special Education and Teaching: 13.1001–13.1099
- Teacher Education & Professional Development, Specific Levels and Methods: 13.1201–13.1299
- Teacher Education & Professional Development, Specific Levels and Methods: 13.1301–13.1399
- Teaching English or French as a Second or Foreign Language: 13.1401–13.1499
- Education, Other: 13.9999

We restrict the data to include only graduates that earned a bachelor's or a master's degree and then sum these university-level counts to the state-by-year level.

[23] Crosswalks can be found here: https://nces.ed.gov/ipeds/cipcode/crosswalk.aspx?y=55

[24] Dan Goldhaber and Roddy Theobald provided valuable guidance here as well.

References

Angrist, J.D., Guryan, J., 2004. Teacher testing, teacher education, and teacher characteristics. Am. Econ. Rev. 94 (2), 241–246.

Angrist, J.D., Guryan, J., 2008. Does teacher testing raise teacher quality? Evidence from state certification requirements. Econ. Educ. Rev. 27 (5), 483–503.

Bowen, Daniel H., Buck, Stuart, Deck, Cary, Mills, Jonathan N., Shuls, James V., 2015. Risky business: an analysis of teacher risk preferences. Educ. Econ. 23 (4), 470–480.

Boyd, Donald, Grossman, Pamela, Lankford, Hamilton, Loeb, Susanna, Wyckoff, James, 2006. How changes in entry requirements alter the teacher workforce and affect student achievement. Educ. Financ. Policy 176–216.

Boyd, Donald, Lankford, Hamilton, Loeb, Susanna, Rockoff, Jonah, Wyckoff, James, 2008. The narrowing gap in New York City teacher qualifications and its implications for student achievement in high-poverty schools. J. Policy Anal. Manag. 27 (4), 793–818.

Britton, Jack, Propper, Carol, 2016. Teacher pay and school productivity: Exploiting wage regulation. J. Public Econ. 133, 75–89.

Chernozhukov, Victor, Fernández-Val, Iván, Melly, Blaise, 2013. Inference on counterfactual distributions. Econometrica 81 (6), 2205–2268.

Chiang, Hanley, Speroni, Cecilia, Herrmann, Mariesa, Hallgren, Kristin, Burkander, Paul, Wellington, Alison, 2017. Evaluation of the Teacher Incentive Fund: Final Report on Implementation and Impacts of Pay-for-Performance across Four Years. NCEE 2018–4004. Center for Education Evaluation and Regional Assistance, National.

Chingos, Matthew M., Peterson, Paul E., 2011. It's easier to pick a good teacher than to train one: Familiar and new results on the correlates of teacher effectiveness. Econ. Educ. Rev. 30 (3), 449–465.

Chingos, Matthew M., West, Martin R., 2012. Do more effective teachers earn more outside the classroom? Educ. Financ. Policy 7 (1), 8–43.

Clotfelter, Charles T., Ladd, Helen F., Vigdor, Jacob L., 2006. Teacher-student matching and the assessment of teacher effectiveness. J. Hum. Resour. 41 (4), 778–820.

Clotfelter, Charles T., Ladd, Helen F., Vigdor, Jacob L., 2007. Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. Econ. Educ. Rev. 26 (6), 673–682.

Clotfelter, Charles, Glennie, Elizabeth, Ladd, Helen, Vigdor, Jacob, 2008. Would higher salaries keep teachers in high-poverty schools? Evidence from a policy intervention in North Carolina. J. Public Econ. 92 (5–6), 1352–1370.

Clotfelter, Charles T., Ladd, Helen F., Vigdor, Jacob L., 2010. Teacher credentials and student achievement in high school a cross-subject analysis with student fixed effects. J. Hum. Resour. 45 (3), 655–681.

Corcoran, Sean P., Evans, William N., Schwab, Robert M., 2004. Changing labor-market opportunities for women and the quality of teachers, 1957–2000. Am. Econ. Rev. 94 (2), 230–235.

Cowan, James, Goldhaber, Dan, Hayes, Kyle, Theobald, Roddy, 2016. Missing elements in the discussion of teacher shortages. Educ. Res. 45 (8), 460–462.

Cullen, Julie Berry, Koedel, Cory, Parsons, Eric, 2019. The compositional effect of rigorous teacher evaluation on workforce quality. Educ. Financ. Policy https://doi.org/10.1162/edfp_a_00292.

Dee, Thomas S., Wyckoff, James, 2015. Incentives, selection, and teacher performance: evidence from IMPACT. J. Policy Anal. Manag. 34 (2), 267–297.

Education Commission of the States. 2016. "Charter Schools - Do teachers in a charter school have to be certified?" Charter Schools - Do teachers in a charter school have to be certified?

Feng, Li, Sass, Tim R., 2017. Teacher quality and teacher mobility. Educ. Financ. Policy 12 (3), 396–418.

Fullan, Michael, 2011. Choosing the wrong drivers for whole system reform. Centre for Strategic Education, Melbourne.

Gershenson, Seth, 2016. Performance standards and employee effort: Evidence from teacher absences. J. Policy Anal. Manag. 35 (3), 615–638.

Goldhaber, Dan, Hansen, Michael, 2010. Using performance on the job to inform teacher tenure decisions. Am. Econ. Rev. 100 (2), 250–255.

Goldhaber, Dan, Walch, Joe, 2014. Gains in teacher quality: Academic capabilities of the US teaching force are on the rise. Educ. Next 14 (1), 38–46.

Goldhaber, Dan, Michael DeArmond, Albert Liu, and Daniel Player. 2007. "Returns to skill and teacher wage premiums: What can we learn by comparing the teacher and private sector labor markets." *School Finance Redesign Project Working Paper* 8.

Goodman-Bacon, A., 2018. Difference-in-differences with Variation in Treatment Timing (No. w25018). Research, National Bureau of Economic.

Gordon, Robert James, Kane, Thomas J., Staiger, Douglas, 2006. Identifying effective teachers using performance on the job. Brookings Institution, Washington, DC.

Hanushek, Eric. 2009. "Teacher Deselection" in Goldhaber, Dan, and Jane Hannaway. 2009. *Creating a New Teaching Profession*. Urban Institute Press. 2100 M Street NW, Washington, DC 20037.

Hoxby, Caroline M., Leigh, Andrew, 2004. Pulled away or pushed out? Explaining the decline of teacher aptitude in the United States. Am. Econ. Rev. 94 (2), 236–240.

Jacob, Brian A., 2013. The effect of employment protection on teacher effort. J. Labor Econ. 31 (4), 727–761.

Jacob, Brian A., Jonah E. Rockoff, Eric S. Taylor, Benjamin Lindy, and Rachel Rosen. 2016. *Teacher applicant hiring and teacher performance: Evidence from DC public schools*. No. w22054. National Bureau of Economic Research.

Johnson, S.M., 2019. Where Teachers Thrive: Organizing Schools for Success. Harvard Education Press, Cambridge, MA.

Klein, Joel, et al., 2010. October 10th. How to Fix our Schools: A Manifesto by Joel Klein. Rhee and other education leaders. *The Washington Post*, Michelle.

Kraft, Matthew A., 2018. In: Hess, R., McShane, M. (Eds.), Federal Efforts to Improve Teacher Quality. Lessons Learned. Harvard Education Press, Bush-Obama School Reform, pp. 69–84.

Kraft, Matthew A., Gilmour, Allison F., 2016. Can principals promote teacher development as evaluators? A case study of principals' views and experiences. Educ. Adm. Q. 52 (5), 711–753.

Lafortune, Julien, Rothstein, Jesse, Schanzenbach, Diane Whitmore, 2018. School finance reform and the distribution of student achievement. Am. Econ. J. Appl. Econ. 10 (2), 1–26.

Lang, Kevin, and Maria Dolores Palacios. 2018. The Determinants of Teachers' Occupational Choice. NBER working paper No. 24883.

Lankford, Hamilton, Loeb, Susanna, McEachin, Andrew, Miller, Luke C., Wyckoff, James, 2014. Who enters teaching? Encouraging evidence that the status of teaching is improving. Educ. Res. 43 (9), 444–453.

Loeb, Susanna, Miller, Luke C., Wyckoff, James, 2015. Performance screens for school improvement: The case of teacher tenure reform in New York City. Educ. Res. 44 (4), 199–212.

Master, Benjamin, Sun, Min, Loeb, Susanna, 2018. Teacher workforce developments: Recent changes in academic competitiveness and job satisfaction of new teachers. Educ. Financ. Policy 13 (3), 310–332.

Nagler, Markus, Piopiunik, Marc, West, Martin R., 2020. Weak markets, strong teachers recession at career start and teacher effectiveness. J. Labor Econ. 38 (2), 453–500.

National Center for Education Statistics (NCES), 2017. Characteristics of Public Elementary and Secondary School Teachers in the United States United States Department of Education.

National Council on Teacher Quality (NCTQ), April 2016. State-by-State Evaluation Timeline Briefs. NCTQ, Washington, DC.

Reback, Randall, Rockoff, Jonah E., Schwartz, Heather L., 2014. Under pressure: job security, resource allocation, and productivity in schools under No Child Left Behind. Am. Econ. J. Econ. Pol. 6 (3), 207–241.

Rockoff, Jonah E., Jacob, Brian A., Kane, Thomas J., Staiger, Douglas O., 2011. Can you recognize an effective teacher when you recruit one? Educ. Financ. Policy 6 (1), 43–74.

Ross, Erin, Walsh, Kate, 2019. State of the States 2019: Teacher and Principal Evaluation Policy. National Council on Teacher Quality, Washington, DC.

Rothstein, Jesse, 2015. Teacher quality policy when supply matters. Am. Econ. Rev. 105 (1), 100–130.

Roy, Andrew Donald, 1951. Some thoughts on the distribution of earnings. Oxf. Econ. Pap. 3 (2), 135–146.

Sartain, Lauren, Steinberg, Matthew P., 2016. Teachers' labor market responses to performance evaluation reform: Experimental evidence from Chicago public schools. J. Hum. Resour. 51 (3), 615–655.

Solon, Gary, Haider, Steven J., Wooldridge, Jeffrey M., 2015. What are we weighting for? J. Hum. Resour. 50 (2), 301–316.

Staiger, Douglas O., Rockoff, Jonah E., 2010. Searching for effective teachers with imperfect information. J. Econ. Perspect. 24 (3), 97–117.

Steele, Jennifer L., Murnane, Richard J., Willett, John B., 2010. Do financial incentives help low-performing schools attract and keep academically talented teachers? Evidence from California. J. Policy Anal. Manag. 29 (3), 451–478.

Steinberg, Matthew P., Donaldson, Morgaen L., 2016. The new educational accountability: understanding the landscape of teacher evaluation in the post-NCLB era. Educ. Financ. Policy 13 (3), 340–359.

Winters, Marcus A., Cowen, Joshua M., 2013a. Who would stay, who would be dismissed? An empirical consideration of value-added teacher retention policies. Educ. Res. 42 (6), 330–337.

Winters, Marcus A., Cowen, Joshua M., 2013b. Would a value-added system of retention improve the distribution of teacher quality? A Simulation of Alternative Policies. J. Policy Anal. Manag. 32 (3), 634–654.