Abstract

This paper employs event history analysis to explore the factors that were associated with the rapid uptake of teacher evaluation reform. We investigate three hypotheses for this rapid adoption: (1) downward diffusion from the federal government through Race to the Top, (2) upward diffusion from large school district policies, and (3) the influence of intermediary organizations. While RTTT clearly played a role in state adoption, our analysis suggests that having a large district implement teacher evaluation reform is the most consistent predictor of state adoption. Intermediary organizations appeared to play a role in the process as well.

Keywords: teacher evaluation, policy implementation, federal policy, school districts, policy innovation and diffusion, state policies, bottom-up reform, education reform

This is the final accepted unformatted version of Bleiberg, J. & **Harbatkin, E.** (2020). Teacher Evaluation Reform: A Convergence of Federal and Local Forces. *Educational Policy*, 34(6), 918 -952. https://doi.org/10.1177/0895904818802105.

Teacher Evaluation Reform:

A Convergence of Federal and Local Forces

In 2009, Washington, D.C. adopted IMPACT, becoming the first state education agency in the nation that required value-added data to count toward teacher evaluations. Within four years, 31 states followed suit, adopting their own teacher evaluation reforms requiring student data components. Partly as a result of these reforms, about 30 percent of teachers–approximately 1 out of every 200 American workers—are now evaluated using student data (Graham, Parmer, Strizek, & Thomas, 2014).[1] This rapid spread of an innovative policy stands out against the traditionally change-resistant backdrop of education institutions (Dimaggio & Powell, 1983).

The goal of this paper is to explore the political dynamics that drove the rapid expansion of teacher evaluation reforms that occurred between 2009 and 2013. This analysis makes three key contributions to the literature. First, the extant literature has found evidence that the American Recovery and Reinvestment Act (ARRA) contributed to the adoption of a broad set of education policies (GAO, 2011; Howell & Magazinnik, 2017; McGuinn, Berger, Stevenson, Hess, & Kelly, 2012). In the wake of the Great Recession, states were faced with unprecedented revenue shortfalls. Congress passed ARRA in 2009 to boost the economy and prevent even more drastic cuts to public programs, which included nearly $5 billion in supplemental education funds. The economic situation was so dire that many states were willing to acquiesce to the education policy preferences of the federal government in return for access to additional funding. Existing research has found evidence that Race to the Top (RTTT) did lead states to adopt education policies favored by the federal government (GAO, 2011; Howell & Magazinnik, 2017; Manna & Ryan, 2011). However, the federal dynamic operated alongside many local ones. Around the same time as RTTT, for example, districts were implementing their own teacher

evaluation reforms in states that had not yet adopted laws (Steinberg & Donaldson, 2016). We pursue a logical extension of this work by focusing more narrowly on teacher evaluation policy.

Second, while a convincing body of research focuses on federal efforts to stimulate state adoption of education reforms via RTTT (Howell & Magazinnik, 2017; McGuinn, 2012), we examine the influence of other activities that occurred in tandem with the RTTT rollout. Our goal is not to suggest that RTTT didn't matter in the wave of state adoptions, but rather to examine the influence of local contexts that may have operated together with RTTT. For example, many districts had already implemented teacher evaluation systems or had plans to do so prior to the RTTT announcement (McGuinn, 2012; Steinberg & Donaldson, 2016). The RTTT effect was in part due to encouraging the adoption of education reforms that were already underway due to nascent local efforts.

Finally, this paper explores the role of intermediary organizations, another actor in the process that led to the rapid wave of state adoptions. Intermediary organizations have the unique capacity to traverse the intergovernmental landscape by influencing each level of government to enact their preferred policies. Organizations such as the Bill and Melinda Gates Foundation have spent tens of millions of dollars to advocate for a narrow set of teacher evaluation policies (Meredith, 2013; Reckhow, 2012). Specifically, we ask: What dynamics are associated with an increased likelihood of state adoption of teacher evaluation policies? To what extent did adoption of teacher evaluation policy diffuse down from the federal government to states? To what extent did adoption of teacher evaluation policy diffuse up from large school districts to states? To what extent was intermediary organization activity associated with state adoption of teacher evaluation policies?

Studies of policy innovation and diffusion have typically investigated the mechanisms driving state education policy adoption or the relationships between the federal and sub-national governments (Berry & Berry, 1990; Shipan & Volden, 2008). To examine the determinants of adoption of teacher evaluation reform, we create a novel dataset by coding district and state teacher evaluation policy implementation and adoption data for 142 of the largest school districts in the country and all 50 states. We aim to contribute to this rich set of policy adoption literature by analyzing education policy diffusion jointly across multiple levels of governance.

This paper proceeds as follows: in the next section, we provide a theoretical framework for the analysis. The third section provides information on the sample and data we used to answer these questions and how we operationalized the measures. The next two section describe our methods and results, while the final section provides a discussion of these results, conclusions, and limitations of the analysis.

Theoretical Framing and Literature Review

Policy innovation and diffusion is a useful framework for understanding why governments choose to adopt a reform. The pattern of policy diffusion often follows an S-shape within a network of governments (e.g., states) over time, in which the policy adoption begins slowly within a small group of early adopters, accelerates rapidly, and then declines after all interested policymakers have adopted the reform (Rogers, 2010). Berry & Berry (1990, 1992) pioneered the event history approach to exploring policy diffusion with a study of state lotteries in which they found that "internal determinants" such as socioeconomic and political factors influenced the adoption of state lottery policies. Subsequent studies found policies could also diffuse vertically from cities to states, but that the relationship depended in part on the level of legislative professionalism in the state legislature and the strength of the policy's advocates.

Cities learn about policies from early adopters, through economic competition with similar cities, and by copying larger cities (Shipan & Volden, 2008).

Studies of the innovation and diffusion of education reform suggest education-related policies diffuse to states through a variety of sources. Mintrom (1997) found that education policy entrepreneurs "venue-shop" by making strategic choices about the level of government at which they will pursue a reform. The presence of intrastate or internal policy networks comprising local policy entrepreneurs was associated with both the consideration and adoption of school choice policies (Mintrom & Vergari, 1998). Other studies investigating the determinants of state adoption of school choice laws found multiple influences, including Republican gubernatorial control, education spending, and the number of private schools in the state (Wong & Langevin, 2006). The number of charter schools in a state was driven by party representation and institutional differences (Zhang & Yang, 2008). Republican-controlled state legislatures are associated with increased likelihood of mayoral control of districts (McGlynn, 2010), and pursuit of RTTT competitive grants was related to adoption the Common Core standards (LaVenia, Cohen-Vogel, & Lang, 2015).

Theoretical framework

Downward vertical diffusion. The downward, or top-down, vertical diffusion hypothesis proposes that the federal government pressures states to adopt a policy, often through funding streams. The federal government can attach conditions to the receipt of education funding. Policy entrepreneurs believe federal forces can be more powerful during economic slowdowns, when states may be more likely to adopt reforms with federal funding is attached (Manna & Ryan 2011). In RTTT in particular, more state budget cuts were associated with higher RTTT scores, suggesting states were responsive to federal education priorities (Manna &

Ryan 2011). Competitive grant programs manufacture a scarce funding stream that may induce states to agree to new policies. After the Great Recession, states found themselves in a dire fiscal situation (McGuinn, 2016). After the RTTT announcement, states replicated policies—including teacher evaluation—from other states that were thought to be favored by the U.S. Department of Education (Meredith, 2013). To improve their case for winning a RTTT grant, some states (e.g., California and Maine) repealed existing laws prohibiting the linkage of student and teacher data for teacher evaluations (GAO 2011). Competition between states, combined with a weakened economy, provided an opportunity for the federal government to pressure states into adopting their preferred education policies.[2]

The competitive grant making process employed by RTTT is an example of an "executive federalism" approach, in which executive powers are strategically employed in the pursuit of state policy reforms (Gais & Fossett, 2005; Howell & Magazinnik, 2017). This approach combines two seemingly contradictory theories of administration. It relies heavily on the voluntary compliance of states to adopt policies. However, the federal government is simultaneously inflexible in the scope of policies that it deems acceptable. Through RTTT and the ESEA waivers, the Obama administration remained tight on goals but loose in how states and districts would meet them (Brown, 2015).

Upward vertical diffusion. Upward policy diffusion—from districts to states— begins with adoption and implementation in a school district. State leaders may recognize improving teacher quality as a goal but lack consensus around the best policy to achieve that goal (Kingdon, 2002). Implementation of a teacher evaluation system in a district generates data for state political leaders to consider. We might expect state education policy makers to monitor the politics of teacher evaluation systems in their state and the effectiveness of those systems toward

improving student outcomes. If the largest school districts in a state have adopted a particular reform, a bottom-up policy diffusion can occur in which state policy makers learn from the local implementation and then adopt state-level reforms(Mintrom & Vergari, 1998; Shipan & Volden, 2006).

After successful district implementation, policy entrepreneurs engage in venue shopping—attempting to scale the reform up to the state level (Mintrom, 1997). Supporters of teacher evaluation reform develop internal policy networks at the state level (Mintrom & Vergari, 1998), and those groups engage local policy actors to build a broad coalition for change. Legislators who support the reform can "borrow strength" from school districts (Manna, 2006). Federal policymakers have similarly leveraged the capacity and license of state education reforms. For example, George H.W. Bush's 1989 Charlottesville education summit of federal policy makers and education reform-minded governors built political support for the passage of Improving America's Schools Act in 1994 (Manna, 2006; Schwartz & Robinson, 2000). Seven years later, the authors of No Child Left Behind strategically included accountability provisions in the law after successful implementation of similar state efforts (Manna, 2006).

A clear example of upward diffusion appears to occur in Connecticut, where New Haven School District in 2010-11 became one of the first school districts in the country to use student data to evaluate teachers (Bailey, 2013). In a 2012 report, the State Board of Education described elements of the New Haven teacher evaluation system and recommended adoption of a statewide system called the System for Evaluation and Development (SEED). Ten school districts across the state participated in a SEED pilot during the 2012-13 school year (NCTQ, 2016). Implementation of SEED rolled out statewide in the 2013-14 school year, and school district adoption of teacher evaluation "softened up" the political ground (Kingdon, 2002). In developing the

state-level policy, policy makers were considering the teacher evaluation systems that districts had already developed and implemented.

The intersections among district implementation, federal influence, and state adoption is not consistently as straightforward as the Connecticut case. In North Carolina, district implementations preceded full-scale state adoption—but received support from the state in doing so. The state offered to make the Education Value-Added Assessment System (EVAAS) available to all districts for free in the 2007-08 school year (LeClaire, 2011), and then piloted the North Carolina Educator Evaluation System (NCEES) the following year (NCTQ, 2016). In the 2010-11 school year, several school districts adopted EVAAS, including two of the three largest districts in the state. North Carolina began requiring districts to use a student growth component in their teacher evaluations the following year. The state's Professional Teaching Standards Commission had conversations about developing the state policy included several members from Wake County (North Carolina Department of Public Instruction, 2009), which was one of the two large districts to implement in 2010-11. Similar to those in Connecticut, North Carolina's policy makers appeared to be considering the implementation of district teacher evaluation policies in their adoption calculus.

However, while Wake and Guilford counties did appear to influence state decisions, federal pressures were also ramped up in that same year. The simultaneous implementation of teacher evaluation in two large school districts and the receipt of a Phase 2 RTTT grant presents a barrier to quantitatively capturing the factors motivating adoption of educator evaluation reforms. To improve its odds at winning an RTTT grant, North Carolina argued in its Phase 2 application that officials installed several new policies and that more would be implemented in the next year (Howell, 2015; Purdue, 2010). The RTTT grant rewarded North Carolina both for teacher evaluation policies

already in place and for expansion plans. It was the joint influence of these factors that resulted in statewide adoption of NCEES in the 2010-11 school year.

Internal determinants. Economic competition between states or election-driven political battles can result in reforms to education policy (Berry, 1994; Berry & Berry, 1990; Wong & Shen, 2002), or states with weak economies may invest in education systems to improve human capital (Doyle, 2006). Other internal state factors relating to education also play a role in a state's propensity to adopt education reform. For example, a state may elect an education-focused governor who pursues broad education reforms (Henig, 2013), or the strengths of teachers unions may hinder state adoption. States with active unions have opposed reforms such as teacher evaluation, and these policies remain unpopular with teachers (Henderson, Peterson, & West, 2016; Moe, 2011).

Intermediary Organizations. Intermediary organizations such as large foundations and other non-governmental organizations may have facilitated the proliferation of teacher evaluation systems across school districts and states. Donations from intermediary organizations have spiked in the past 20 years (Reckhow & Snyder, 2014). In the case of teacher evaluation reform, foundations strategically built partnerships with cities and states around the country, making long-term investments moved toward achieving the foundations' policy objectives (Reckhow, 2012). Although education foundations have different leadership and missions, their donations tend to converge over time to the same group of organizations, which can amplify the influence of their grants (Reckhow & Snyder, 2014). The Gates Foundation and other intermediary organizations have advocated for the adoption of teacher evaluation systems, providing millions of dollars in grants and technical support to states that reformed their teacher evaluation laws (Meredith, 2013).

An example of the intermediary organization role played out in Pennsylvania, which launched a five-district teacher evaluation pilot with support from two Gates Foundation grants in 2010 and 2013. The pilot then expanded to include 120 school districts the following year (NCTQ 2016). In total, the Gates Foundation provided more than $1 million to the state to support its teacher evaluation efforts, and had previously awarded $40 million to Pittsburgh—one of the five initial pilot districts—for participating in the Measures of Effective Teaching (MET) project (Gates Foundation, 2017). The intermediary role operated together with the federal and district roles to support state adoption.

There are several political, financial, and social dynamics that influence how state actors approach policy making. States respond to the federal government, which strategically uses grants to achieve reform objectives. In addition, state policy makers are also politicians and are keenly aware of the developments in their own state. The successful implementation of policies in large school districts within a state is therefore instructive as policy makers weigh policy alternatives. Recently, intermediary organizations have occupied a growing role in reform processes primarily through their capacity to offer financial support in return for reform. We explore the role of each of those three mechanisms in the diffusion of teacher evaluation policy reform.

Sample, Data, and Measures

Sample

Our analysis draws from two linked samples: a sample of all 50 states plus Washington, D.C., and a sample of large school districts. The event history analysis drops Washington, D.C., and Hawaii because both contain single school districts, and excludes Nebraska because the state

has a unicameral legislature and cannot be included in models that include state legislative variables.[3]

At the district level, our sample consists of the 100 largest school districts in the country, supplemented to include at least two from each state for a total of 142 districts. The 100 largest school districts in the country and the largest in each state were identified by the National Council on Teacher Quality web-based teacher contract database (NCTQ, 2016). For states that had only one district in the NCTQ database, we identify the second-largest district using average daily membership figures from the Common Core of Data (NCES, 2014). The average school district in the sample contains 4,540 teachers, about 75,000 students, is located in a county with a 2013 per capita income of about $48,600, and educates about 24 percent of the students in its home state. The school district sample is not intended to be representative of all U.S. school districts, but rather because we assume that the largest school districts in a state is likely to have an outsized influence on state policy (Mintrom & Vergari, 1998).

We observe policies in these states and districts from 2009 through 2015. We start in 2009 because it is the first year in which states began adopting teacher evaluation reform, and continue through 2015 because Congress announced it reached a deal to pass the every Student Succeeds Act (ESSA) (Severns, 2015), which replaced, reauthorized, or reorganized RTTT initiatives. As a result, states that adopted teacher evaluation systems in that year likely were responding to a new set of political dynamics that are beyond the scope of our analysis.

Data and Measures

The dataset for this analysis draws from multiple sources, including our coding of state and district policies. The time period spans 2009 through 2015, when the rapid uptake of teacher evaluation reform occurred. The dependent variable is a binary indicator denoting whether the

state adopted teacher evaluation reform with a student data component. The independent variables of interest each align with one of the four hypothesized mechanisms of policy diffusion:

- Diffusion down: Race to the Top wins and losses
- Diffusion up: Large district implementation of teacher evaluation
- Intermediary organizations: Gates grant award
- Internal determinants: State political, economic, and educational characteristics

We describe each of these variables in detail below.

Dependent variable. The dependent variable in the event history analysis is a binary indicator for whether or not the state adopted teacher evaluation reform in that year. In this analysis, we focus specifically on adoption of teacher evaluation systems with a student data component weighted as part of teacher evaluation scores.[4] We focus on this narrow definition of teacher evaluation reform for two key reasons. First, accurately testing our theory that local forces worked in tandem with RTTT to influence state adoption suggests a definition of teacher evaluation that aligns with a high RTTT application score, a large part of which is tied to adoption of teacher evaluation policies drawing from student data. The largest component of the RTTT scoring rubric is the Great Teachers and Leaders score, which rewarded states for having a reform plan that included rigorous evaluation systems differentiating teacher and principal effectiveness by levels, annual teacher and principal evaluations, and use of those annual evaluations in decisions around compensation, promotion, retention, and tenure (U.S. Department of Education, 2009).[5,6] Second, this operationalization reflects salient differences in the actual features of teacher evaluation policies, which differed considerably across states (Kraft & Gilmour, 2016) through inclusion of different components (i.e., observations, Student

Learning Objectives (SLOs), VAM, surveys, professional conduct measures, and schoolwide achievement). Some states and district made modest changes to their previous systems (e.g., SLOs). In their execution, these systems were qualitatively similar to previous policies. Including a student data component represented an innovation and a divergence from the "business as usual" of teacher evaluation.

The primary goal of this analysis is to explore the forces that influence state policy adoption. However, focusing exclusively on state adoption misses some of the nuance that comes with state policy making—specifically, states may adopt policies that they never fully implement, or may not implement a policy until years after initial adoption. For this reason, we conduct a supplementary analysis in which the dependent variable is the gap between adoption and implementation year. In this supplementary analysis, we explore the influence of our independent variables of interest on the length of that gap, with the goal of better understanding not just the determinants driving policy adoption—but also the determinants associated with more immediate implementation of the policy that has been adopted.

Independent variables. In alignment with our theory that RTTT, district implementation, and intermediary organizations all played a role in state adoption of teacher evaluation reform, we investigated three key measures of interest: (1) RTTT status by award phase, (2) school district implementation of teacher evaluation within a state, and (3) state and local teacher evaluation grants from the Gates Foundation.

Race to the Top. We follow the coding strategy of Howell & Magazinnik (2017) by creating separate binary variables for RTTT Phase 1 and 2 winners and losers, respectively, and excluding Phase 3 recipients. Each win and loss variable takes the value of 1 in the year of award and in each year thereafter. We do not include a variable for Phase 1 winners because the two

states that won Phase 1 awards (Delaware and Tennessee) both adopted teacher evaluation systems prior to receiving the federal award. It is therefore not possible to code the *RTTT Phase 1 win* variable in the same way as the other RTTT variables given our event history framework because the "failure" in these states occurs prior to Race to the Top. We focus only on the first two phases because Phase 3 recipients were not subject to the same pressures to adopt new policies favored by RTTT. As Howell & Magazinnik (2017) described, only losing Phase 2 finalists were invited to apply, and the application required states to select a subset of activities from their losing Phase 2 applications rather than being scored on all RTTT components. Additionally, many of the performance measures on which states were evaluated were left to state discretion.

District teacher evaluation implementation. While the dependent variable represents whether or not a state adopted teacher evaluation, we operationalize the district variable as implementation rather than adoption for two key reasons. First, the theory of action around upward diffusion that states learn from district policies requires that districts have implemented the policy being considered. The policy learning that takes place at the state level requires implementation of the policy in question—not just adoption. Second, a diffusion-down scenario would yield state adoption followed by district implementation, and coding accordingly allows us to be more confident about the order of events given the limitations associated with discrete time. Finally, the available online resources on district policy are more reliable for implementation data than adoption data because readily available resources vary so much from district to district. After coding all 142 districts in the district sample, we collapse district implementation to a single binary measure that takes a value of 1 if any district in a state had implemented by that year.[7]

To code district policy, we conducted a document review of collective bargaining agreements (CBAs), human resource materials, and other publicly available documents related to teacher evaluation policy within districts. We drew these documents from the NCTQ teacher contract database and internet searches. Based on the review, we coded districts as having implemented teacher evaluation if the evaluation system was fully implemented and included a measure derived from a standardized test (e.g., proficiency rates, student growth measures, teacher value added). In order for the evaluation system to count in our model, we require that the student data have some weight toward the teacher evaluation. Systems with optional inclusion of data, a "significant factor" requirement, SLOs, or formative assessments are not counted as teacher evaluation unless the system also included the student data component as described above. The inclusion of sanctions or rewards does not determine if the system counts as teacher evaluation.[8]

We do not count small district pilots but also do not require universal implementation across district schools and grades to code a district as having implemented. The definition of a teacher evaluation system "pilot" varied widely across districts and states. When a system was described as a pilot, we examined each case and made a coding decision based on the scope of implementation. If a preponderance of teachers in a district or school districts in a state were using a student data component in their evaluation system, then we count the district as having implemented.

Intermediary organizations. In an effort to operationalize intermediary organizations, we create two binary variables: one indicating whether the state won a Bill and Melinda Gates Foundation grant related to teacher evaluation and one indicating whether a district in the state won a grant focused on teacher evaluation at any time prior to the observed year. These variables follow

the same format as the RTTT and district implementation variables, taking a value of 1 in the first year of the grant and for each subsequent year the state remains in the analysis. These grants are not representative of all intermediary organization involvement or even all education-related Gates Foundation grants during those years—and by only counting grants specifically related to teacher evaluation programs, this variable provides a conservative measure of Gates involvement. We focus on the Gates Foundation both because there were one of the most active boosters of teacher evaluation and because education funding converged around their priorities (Reckhow & Snyder, 2014). Despite the limitations associated with focusing on a single foundation, we believe that capturing a subset of grants from one of the largest and most active foundations in education reform does provide a valuable measure of intermediary organization involvement in teacher evaluation policy.

To generate this variable, we conducted a search on the Gates Foundation website of all grants related to primary and secondary education in the United States from 2009 through 2016 (Gates Foundation, 2017). From those about 1,200 grants, we use program summaries to determine whether the grant was related to teacher evaluation and when necessary conduct internet searches for additional available information about the program. These Gates Foundation database also includes grant amounts, but we dichotomize the measure because in many cases the grant amounts were based on average daily attendance and were not a judgment on the teacher evaluation program size or quality. For example, the largest grants went toward Measures of Effective Teaching (MET) programs, for which grant sizes were largely determined by the number of students in the school district. Grants were awarded by Gates to 11 different states as well. Those 11 states ranged from one to 11 grants during the seven-year time period, with five of them

winning in multiple years. Thirty local grants were awarded to 17 school districts in 15 states. Those 17 districts ranged from one to four grants during the study period.

Internal determinants. We control for state educational, political, and economic characteristics. Political covariates include whether the state's governor was a member of the National Governors Association (NGA) education committee in that year and the governor's political party (NGA, 2017)[9], state legislative professionalism scores as calculated by Bowen and Greene (2016), a binary variable taking the value of 1 in the years a state has a Democratic governor, and a variable denoting the state's Democratic legislature percentage (The Council of State Governments., 2015). The NGA variable is intended to capture both a governor's propensity to adopt education reform and for policy learning that may occur across states but not necessarily as a result of geographic proximity. Membership in the NGA committee may reflect a governor's early interest in promoting and adopting state education reform regardless of RTTT receipt or district policies, and it may also provide opportunities for policy learning by offering a clearinghouse of information about potential policies (Shipan & Volden, 2012). We include the Democratic governor and legislature variables to account for any partisan influence on policy adoption. While teacher evaluation has not emerged as a strongly partisan issue on either side of the aisle, a subset of research investigating the determinants of state education policy did find a significant relationship between state politics and policy adoption. For example, in an event history analysis exploring the effect of state characteristics on charter school legislation, Hassel (Hassel, 1999) found that having a Republican governor was a significant predictor of charter policy adoption. In another study of charter school policy, (Renzulli & Roscigno, 2005) found evidence that an increase in Republican legislators was associated with higher odds of adopting a strong charter law, which the authors defined as laws that reduce restrictions and streamline

processes for establishing charter schools—but did not find that having a Republican governor mattered.

Educational covariates include a measure of the state's most recent performance on the National Assessment of Educational Progress (NAEP) (NCES, 2016)[10], state per pupil expenditures (U.S. Census 2017), logged state K-12 enrollment (NCES, 2017), and the percentage of teachers in a state who are members of a teachers union. The NAEP variable is intended to capture the state's educational quality relative to the rest of the country. We might expect states with lower NAEP scores to have a stronger incentive to implement educational reforms in an effort to catch up to higher performing states. The union membership variable represents the strength and breadth of teachers unions in that state (Winkler, Scull, & Zeehandelaar, 2012), although it is limited by its time invariance. Ideally, we would include a teachers' union membership variable that is updated each year. However, no such data are available during the study period and historical data show that while private sector union membership has steadily declined in recent years, public sector membership has remained relatively stable since the early 1980s (U.S. Bureau of Labor Statistics, 2016). Past research has found strong unions can act as a barrier to implementation of education reforms (Moe, 2011), suggesting states with more union strength may be less likely to adopt teacher evaluation. Per pupil expenditures is intended to control for the state's existing level of educational spending, and K-12 enrollment captures the magnitude of the primary and secondary schooling needs. Finally, we include state GDP and unemployment rate to capture the state's economic circumstances, which may play a role in propensity to adopt reforms.

Methods

We aim to isolate the roles of the three hypothesized mechanisms of diffusion using a discrete-time hazard model of state adoption to model the overall baseline hazard with a time trend to capture changes over time. We are limited in our analysis by a short observation time period, interval censoring resulting from unknown order of events when states adopted in the same year as districts implemented, and a small number of failures (*N*=31). Given these limitations and the resulting limited power to isolate the separate contributions of our hypothesized mechanisms, our goal is to first ask whether there are any associations between state adoption and the variables of interest. As such, we begin by estimating naïve logit-transformed discrete time hazard functions with state policy adoption on the left side and each of the constructs of interest on the right with three separate models: district implementation, Gates awards (state and local), and RTTT (wins and losses by round). We then add a set of covariates reflecting the state's educational, political, and economic characteristics as described above. This first set of models pools all years of data from 2009 through 2015 to estimate a mean baseline hazard rate, taking the form

$$h_i(t) = \exp\left[\beta_0 + \beta_1 Z_{it} + \alpha \boldsymbol{S'}_{i(t)} + \gamma \boldsymbol{P'}_{it} + \delta \boldsymbol{E'}_{it} + e_{it}\right]$$

where $h_i(t)$ is the log-transformed discrete time hazard function for state *i*; Z is a switch is replaced by the RTTT vector in Model 3, the district implementation indicator in Model 6, and the Gates award vector in Model 9; *S* is a vector of time-varying and non-varying state educational variables; *P* is a vector of time-varying state political variables, *E* is a vector of time-varying state economic variables, and *e* is an idiosyncratic error term.

After exploring these simple associations, we estimate a fully specified model with all three mechanisms of interest and a set of covariates, taking the form

$$h_i(t) = \exp\left[\beta_0 + \beta_1 RTTT'_{it} + \beta_2 DistImplement_{it} + \beta_3 Gates'_{it} + \alpha S'_{i(t)} + \gamma P'_{it} + \delta E'_{it} + e_{it}\right]$$

where *RTTT* represents the vector of RTTT variables (Phase 1 loss, Phase 2 win, and Phase 2 loss); *DistImplement* is an indicator denoting whether state *i* has a district that implemented by year *t*, *Gates* represents two indicators in which the first denotes whether the state received a Gates grant in that year and the second denotes whether a district in the state received a Gates grant, and the rest of the model follows the same structure as the associational model above. Finally, we estimate all four covariate-adjusted models with a nonlinear year trend to account for the changing rate of state adoption over time by including the year and the year squared.[11]

To explore the determinants of the gap between adoption and implementation, we limit the analytic sample to states that adopted and then use ordinary least squares to regress the gap length on each of the three mechanisms of interest, separately and then in a single equation taking the form

$$GapLength_i = \beta_0 + \beta_1 RTTT_i + \beta_2 DistImplement_i + \beta_3 Gates_i + e_i$$

where the values of the *RTTT*, *DistImplement*, and *Gates* variables represent the values in the year of state adoption. In this secondary analysis, the analytic sample includes only the states that adopt and we observe each state once rather than over multiple years (n=30) because the outcome is the length of the gap rather than a dichotomous failure variable as in the case of the event history analysis. For this reason, we do not include internal determinants covariates because while implementation of an adopted policy may depend partly on the value of the covariates in the year of implementation, failure to implement may result of the value of those covariates in the years when implementation did not occur.

Results

We begin with a descriptive analysis to explore the trends in adoption of teacher evaluation policies—individually and then in relation to RTTT, school district implementation, and intermediary organizations. The event history analysis then aims to hone in on more precise estimates of the influence of these three factors on state adoption.

Documenting Adoption

States began passing teacher evaluation reform in 2009, with Washington, D.C., Louisiana, and Tennessee as the first to adopt. The following year, eight states adopted reforms, propagating an upsurge in state adoptions that would last the next three years. Meanwhile, early district implementation began slowly increasing in 2010 and spiked in 2012 and 2013. The right skew of state adoption alongside the left skew of first district implementation in Figure 1 suggest a traditional top-down mechanism because more states than districts adopted in the early years and district implementation escalated as state adoptions subsided. Indeed, the 2010 spike in state adoption aligned with the rollout and awards for the first two rounds of RTTT funding—although the cluster of 2009 and 2010 district implementers suggest another mechanism may have been at play in driving states to adopt teacher evaluation reforms because many of these districts implemented reforms before their states required it.

To better understand the federal influence, we investigate the intersection between RTTT award year and state adoption year. The first round of funding for RTTT was announced in 2009. At that time, the U.S. Department of Education published its notice of final priorities, which valued the Great Teachers and Leaders scale more highly than any other scales in the competitive grant. The first round was awarded in March 2010, the second in August 2010, and the third in December 2011. Descriptively, there was a close relationship between winning

RTTT funds and adopting teacher evaluation reform—both Phase 1 winners and 9 of 10 Phase II winners adopted teacher evaluation reform. But the potential for RTTT funds does not appear to tell the full story: one Phase 2 winner (Maryland) won a Phase 2 award but never adopted reform that falls under our definition, and two states (Alaska and North Dakota) adopted despite not applying for RTTT in either round. Additionally, 7 of the 13 states that lost Phase 1 and 2 adopted after the Phase 2 awards were announced, and four of those did not receive Phase 3 awards. For these states, as well as the two Phase 2 winners that adopted teacher evaluation policies two years after receiving the RTTT award, something outside of RTTT may have been contributing to that adoption.

We turn next to our second hypothesis, which posits that states adopted teacher evaluation policies in response to large district implementations. On the surface, the district influence is not as obvious as the federal influence: states overwhelmingly adopted teacher evaluation reform before the first school district in their state implemented a program (Table A-1). Seven states did not adopt despite having a district implement. However, five states (Connecticut, Georgia, North Carolina, Pennsylvania, and Texas) had districts implement first, and three (Colorado, Florida, and D.C.) adopted in the same year as their first district implemented. Our goal is to better understand the extent to which district implementation in those eight states may have influenced state adoption—either on its own or in concert with RTTT and intermediary organizations such as Gates.

Bringing together the first two hypotheses, Figure 2 shows how RTTT wins correlated with state passage/district implementation directionality. While RTTT winners adopted prior to large district implementations in most cases, a subset of RTTT winner and loser states adopted subsequent to large district implementations. The two Phase 1 winners adopted before their large

districts implemented, but there was less consistency among Phase 2 applicants and states that did not apply for RTTT funds at all. The darkest-shaded bar segments in suggest district implementation may have played a role in state decisions to adopt. Among the 36 Phase 2 applicants, two winners and two losers adopted following a large district implementation in their respective states.

Finally, we explore our third hypothesis—that intermediary organizations such as the Gates Foundation played a role in state adoption—by exploring the association between Gates awards and state adoption. Figure 3 shows two clear patterns. First, the Gates Foundation operated at the forefront of the teacher evaluation reform trend. Gates awarded five grants to states and six to districts in 2009—all in states that would adopt teacher evaluation reform. Second, states that received awards themselves or that had large districts receive awards overwhelmingly adopted teacher evaluation reform in the year or years following the award. This pattern is illustrated in Figure 3, where markers above the 45-degree diagonal denote states that adopted following their first Gates award, markers on the diagonal represent states that adopted in the same year as the first award, and those below the diagonal are states that had already adopted by the time they won their first Gates award.[12]

Event History Analysis

We begin by examining Kaplan-Meier curves to explore the failure rate starting at our baseline year of 2008. As shown in Figure 4(a), the failure rate steadily climbs from 2009 through 2012 and then tapers off in later years as fewer states adopt new policies. The earliest adopters were largely RTTT Phase 1 and then Phase 2 winner states, as shown in Figure A-5(b), providing some visual evidence that the federal government influenced early adoptions of teacher evaluation reform. The lower curves show that RTTT losers continued to adopt after the

promise of federal funds had dissolved, and states that did not apply for RTTT funds began adopting in 2013—after all three rounds of awards were granted, again suggesting mechanisms of policy diffusion outside of the federal influence. Figure A-5(c) and (d) show that states with large district implementers and Gates grants, respectively, began adopting teacher evaluation policies at higher rates than states without district implementers and Gates grants in 2011—the year after RTTT Phase 2 grants were awarded—and 2012. To parse out these three separate influences, we turn next to the discrete time hazard regression results. These results are organized by our three diffusion hypotheses. Under each hypothesis, we begin by describing results from the naïve model, which regresses state adoption on just the parameter of interest (Race to the Top wins and losses in Model 1, district implementation in Model 4, and Gates awards in Model 7). To each of those separate naïve models, we then add internal determinants covariates (Models 2, 5, and 8), and then a nonlinear time trend (Models 3, 6, and 9). These results are shown in Table 1. Then we describe results from the full model (Table 2), which includes the variables representing all three mechanisms of interest together and internal determinants covariates in Model 10, and finally adds a nonlinear time trend in Model 11 for the fully specified model.

Hypothesis 1: Top-down diffusion. In alignment with the conventional wisdom around top-down diffusion, our naive models (Table 1) find a positive association between winning RTTT 2 and adopting a teacher evaluation system in Model 2, which focuses only on RTTT and controls for internal determinants. Most states that won RTTT 2 also applied for and lost RTTT 1. For these states, the federal effect of winning a Phase 2 award is the linear combination of the RTTT 1 loss variable and the RTTT 2 win variable, which suggests these states were 7.3 (naïve model) to 12.3 (model with covariates) times more likely to adopt in RTTT years or later ($p<.10$). However, when

we add the time trend to the RTTT-focused models (Model 3), the joint effect size diminishes and is no longer significant (p=.12). We believe there are two potential reasons for the reduced effect size and significance: first, the short timeframe during which most of the state adoptions occurred limits the power in this analysis. It is possible that we can no longer detect an effect with the time trend included. Second, it is possible that late adopters that won Phase 2 awards were adopting in response to other influences alongside states that did not win awards. In this scenario, it is possible that RTTT 2 winners that adopted in later years would have done so with or without the federal influence. When we introduce district implementation and Gates grants into the models (Table 2), we again do not detect a significant increase in the probability of adoption when the time trend is included (Model 11). However, in the model without a time trend (Model 10), losing Phase 1 on its own appears to be associated with a greater likelihood of adoption, and the joint effect of losing Phase 1 and winning Phase 2 is again positive and significant—suggesting a federal influence as RTTT rolled out. Meanwhile, losing RTTT 2 is associated with a decreased likelihood of adoption in the model with the time trend. This negative association likely appears because states that did not adopt in anticipation of Round 2 or immediately after were unlikely to adopt at all. The federal influence therefore appears to be limited to the time period in which the awards are being offered. Additionally, the states that lost Phase 2 may have lost points specifically because their teacher evaluation plans did not closely align with federal priorities. Given the limited power of this analysis, we consider these results together to suggest the federal government likely played a role in the rapid adoption of state teacher evaluation systems, but the federal influence operated in tandem with other state-level dynamics rather than in a vacuum.

Hypothesis 2: Bottom-up diffusion. Three of five models exploring the district role support our hypothesis that having a large district implement is associated with higher odds of

adopting a state policy (Table 1 and 2). The full model including the time trend (Model 11) suggests the odds of adoption are 6.0 times higher in states in which a large district implemented, controlling for Race to the Top, Gates grants, and all other covariates. While including the time trend reduces the RTTT estimate, it does the opposite for district implementation. This pattern suggests that although the early state adopters preceded most district implementations, the districts that implemented in later years seemed to consistently influence their home states—whereas the RTTT role mostly operated during the RTTT award years.

Hypothesis 3: Intermediary organizations. Gates awards to districts are associated with state adoption across all models. However, the magnitude of the district award estimate decreases when RTTT and district implementation are introduced into the model (Models 10 and 11), suggesting that these awards may have had both a direct and indirect influence on state adoption. Awards directly to states are associated with state adoption in the models that omitted RTTT and district implementation (Models 7 and 8), but we do not detect an effect of state awards in the fully specified model (Model 11). If intermediary organizations such as Gates were playing a role in state adoption through direct grants to states, they appeared to be doing so through other mechanisms; for example, states may have received Gates grants for teacher evaluation systems that bolstered their RTTT applications. The decreased estimate on district awards in Model 11 suggests the role of grants to district may have followed a similar pattern: districts received Gates grants that allowed them to implement a district program that was ultimately replicated at the state level. Given limited power in this analysis, we do not introduce interactions to further explore these mechanisms.

Internal determinants. In all of our models, we found higher union membership was associated with decreased likelihood of adoption, suggesting unions may have been successfully

pushing back against teacher evaluation policies. Legislative professionalism was associated with an increased probability of adoption in 5 of the 7 models that included it. None of our other internal determinants covariates consistently reached conventional levels of statistical significance.

Gap analysis

In the supplementary analysis exploring the association between the three mechanisms of interest and the length the adoption-implementation gap, we find a marginally significant association between losing RTTT 1 and implementing teacher evaluation. The adoption-implementation gap in states that lost RTTT 1 was 1.6 to 1.7 years smaller than the gap for states that adopted but did not apply for RTTT funds (Table 3). These results suggest losing Phase 1 may have led states to accelerate the implementation of policies likely to bolster their Phase 2 scores. We do not find evidence that RTTT 2 hastened implementation. States may have been more responsive to implementing federally favored policies in the early stages of RTTT, before they observed that the federal government did not penalize states for failing to implement. We also do not find that district implementation or Gates grants are associated with a significant change in the speed of implementation following adoption.

Discussion

The federal government through Race to the Top clearly played a role in the rapid uptake of teacher evaluation reform by states. However, Race to the Top does not appear to tell the whole story. This finding is consistent with findings by (Howell & Magazinnik, 2017), who find that the federal government influenced states through both direct and indirect pathways. Our analysis suggests that having a large district implement teacher evaluation reform is the most

consistent predictor of state adoption of reform. Intermediary organizations appeared to play a role in the process; however, their involvement may operate through other mechanisms including Race to the Top and large district implementation. Given the limited sample size and short timeframe, we do not have sufficient power to test the hypothesis that awards from intermediary organizations such as the Gates Foundation moderated the uptake of state adoption through these mechanisms.

A limitation of these findings is that our definition of teacher evaluation reform for the collapses a multifaceted measure into a binary variable. Contemporary teacher evaluation systems include three types of components: student data measures, observations, and student feedback. Each component has a weight and may trigger a sanction or reward. In our ideal analysis, we would be able to operationalize this measure in greater detail but the short time period of study combined with the wide variation in state- and district-level policies—and the way they are described in policy documents[13]—constrains our ability to do so. This limitation would bias our results only if the direction of imprecision in the dichotomous variable were correlated with one of our independent variables of interest, which we do not expect to be the case. Instead, this imprecision should random and not systematic.

Second, while our district-level dataset includes 142 large school districts representing the largest districts in each state, it is plausible that small- and medium-sized district policy also contributed to state adoption and our analysis would not capture that effect. These limitations are a particular concern in states like New Jersey that primarily have small and medium school districts. Additionally, the data presented in this paper are specific to the policy context of the day. ESSA could complicate the conclusions we draw here.

Some of these limitations reduce the variation we are able to exploit in testing the four mechanisms in our theory of action. For example, the measure of the role of upward diffusion includes only the influence of the districts in our dataset, and our estimate would therefore understate the role of district implementation if smaller districts were driving state adoption. In evaluating the federal role to test the downward diffusion hypothesis, we are somewhat limited by the binary measures of RTTT wins and losses. These measures represent the extent to which state applications aligned with federal priorities including student data-oriented teacher evaluation—they do not capture the extent to which states actually implemented the proposed reforms. Our estimate of downward diffusion therefore captures proposed state reforms, which may not necessarily match actual reforms (Pressman & Wildavsky, 1984). In examining the role of intermediary organizations, we chose to focus only on Gates award dollars. This narrow focus captures only one funding stream from intermediary organizations, and the estimate will understate the role of these organizations to the extent that others incentivized states and districts to adopt similar reforms. Finally, in exploring the influence of internal state determinants, the measure of union membership is time-invariant and therefore does not capture changes in union membership that may arise in the presence of state reforms. While the short time panel diminishes the possibility of bias arising from year-to-year changes, the estimated role of union membership is likely to be attenuated due to the noisiness of the available measure.

Conclusion

The genesis of state adoptions occurred around the same time as the rollout of RTTT, which encouraged teacher evaluation reform through its competitive grant process. Additionally, a 2011 U.S. Government Accountability Office (GAO) noted that state officials said they changed teacher evaluation laws specifically in response to the RTTT program (GAO, 2011;

McGuinn, 2012). These two reasons together likely drove the public narrative that the swift diffusion of teacher evaluation reform stemmed from the federal influence. However, the patterns of state adoption in relation to RTTT award and district implementation suggest RTTT did not singlehandedly lead to these reforms. Even after the third and final RTTT funding round, states continued to pass teacher new evaluation reform laws and in some cases those laws were passed after the state's largest districts had already implemented a program. Meanwhile, naïve models suggest that state Gates grants had a moderating influence on states' likelihood of adoption, indicating intermediary organizations played a role as well.

These findings suggest the relationships between different levels of government are dynamic and multifaceted (Marsh & Wohlstetter, 2013). Our findings provide a useful framework for understanding how the passage of ESSA may alter the relationships between the federal government and states. Some researchers expect that ESSA represents a shift that will relegate education policy making powers from the federal government to the states (Saultz, Fusarelli, & McEachin, 2017; Weiss & McGuinn, 2016). According to this strain of research, the federal government gained power through NCLB, RTTT, and the ESEA waivers, and has conceded some of that power with ESSA—by allowing states greater discretion in writing school accountability plans and reining in the Education Secretary's waiver authority. Our results suggest that while federal pressure was a salient factor in state education policy making, the actions of school districts on the ground mattered, too. We propose that the new federal policies do not represent a dramatic shift from the status quo, and will not change the degree to which local factors influenced state adoption of education policies.

Our analyses also highlight the role of policy learning in the diffusion of teacher evaluation policies from large school districts up to states. To some extent, these findings

reframe the role of school district leaders. Implementation of large policies influences not only the students and teachers in their own districts, but also the choices of state policy makers. This power presents a possible channel for district leaders that want to assert a larger role for school districts in the national education reform conversation.

Future research ought to investigate the role of intermediary organizations. Education reform focused organization have pursued their goals through numerous channels beyond the awarding of grants to states in districts. (Meredith, 2013) found that the Gates Foundation provided consulting services to states that wanted to improve the their RTTT applications. There are likely other financial and political channels through which intermediary organizations pursued their goals. Better understanding of each of the activities and their influence on policy adoption would help to expand upon these results. Finally, the present study focuses on determinants of state adoption and does not delve into changes to policy implementation at the state and district levels over time. Future research might examine the extent to which the policies that states adopted were actually implemented and the determinants of implementation. Future research could also consider how district policies evolved under ESSA.

---

[1] There are 3.1 million public school primary and secondary teachers (NCES, 2015) in 2017 and 160 million member of the civilian labor force during that same time (BLS, 2017). 29.4 percent of teachers in large districts are evaluated using Value Added Measure or Student Growth Percentiles. A similar percentage of teacher in 2011-2012 reported that their evaluations included student data in the restricted use Schools and Staffing Survey (Graham et al., 2014).

[2] We choose to not account for whether states received ESEA waivers in our empirical analysis. First, Wong (2015) found that the Obama administration encountered difficulties in using the waivers to promote their preferred teacher and principal evaluation reforms. This suggests that the waivers did not have an effect on state adoption of teacher evaluation systems. Descriptively, we observe few states in our sample to have adopted teacher evaluation systems after the announcement of the waivers. The Department of Education made the inclusion of a student growth component in the teacher evaluation system a condition receiving of receiving a waiver for 6 states (Alabama, Iowa, Kansas, South Dakota, Texas, and Washington; (NCTQ, 2016). To date, none of those states have implemented teacher evaluation system with a student data component.

[3] We also estimate the models including Nebraska and get very similar results. Those results are available upon request.

[4] The student data component only needed to apply for teachers of tested subjects in tested grades under No Child Left Behind (NCLB).

[5] The Great Teachers and Leaders score was worth 138 of the 500 possible points on the RTTT application, and states could earn another 47 points for the Data Systems to Support Instruction score, which called for having implemented a statewide longitudinal data system, using data to improve instruction, and accessing and using state data.

[6] Our coding strategy also parallels a survey item from the Teacher Follow-up Survey which was part of the 2011 Schools and Staffing Survey. Teachers who reported participating in a formal evaluation procedure were asked "Are student test score outcomes or test score growth included as an evaluation criterion in your FORMAL evaluation this school year?" (Graham et al., 2014)

[7] We also ran models with (a) the number of districts that had implemented in a state within a year, and (b) the cumulative number of districts that had implemented by that year. All three models yielded similar results; we choose to present the binary indicator to allow for a more easily interpretable story. We do not have implementation data on all districts within a state and the districts we do have represent a wide variety of proportions of state enrollment.

[8] Prior to 2009, school districts did use student data for merit pay or involuntary dismissal (NCTQ, 2010; Springer & Gardner, 2010). However, none of the districts in our sample used student data for teacher evaluation purposes. This finding is consistent with Steinberg & Donaldson (2016).

[9] Some governors had discontinuous membership in the NGA committee. In these cases, we coded the state as a 1 in each year from the first to the last year of a governor's membership. Some of these cases came in states with biennial legislatures, suggesting governors may have been attending only in years when they had an opportunity to influence introduction of new policies. Other governors may have simply missed an NGA meeting and therefore missed the opportunity to participate in the education committee. Data were collected from the current NGA Committee Membership list and archived websites through the Internet Archive.

[10] The NAEP variable is a standardized composite of the state's eighth-grade math and reading NAEP scale scores. Because the NAEP is administered every two years in odd years only, we assign the standardized 2009 measure in 2009 and 2010, the 2011 measure in 2011 and 2012, and so on.

[11] We also estimated models with year fixed effects to allow the constant to shift in each year. The results are substantively similar. We chose to show the year trend results because no states adopted in 2014 and 2015, limiting our ability to estimate year fixed effects in the final two years of our analysis. The fixed effects results are available upon request.

[12] A table showing each state's adoption year, each state's earliest district implementation year, Gates grants, and RTTT status can be found in Tables A-2 and A-3.

[13] In coding district documents such as collective bargaining agreements, we can only confirm that an agreement around teacher evaluation was reached, and not that the teacher evaluation policy was actually carried out as described. Teacher evaluation policies have required observations of teachers for decades, but many of these mandated activities do not actually occur (Weisberg et al., 2009), so we have no way of knowing whether the district implementations we identified actually took effect.