# Investing in Development:  Six High-Performing, High-Poverty Schools Implement Massachusetts' Teacher Evaluation Policy

Stefanie K. Reinhorn, Independent Consultant

Susan Moore Johnson, Harvard Graduate School of Education

Nicole S. Simon, John Jay College of Criminal Justice, CUNY

April 20, 2016

For well over a century, U.S. school administrators have been expected to evaluate teachers (Cubberly, 1915), but until very recently evaluation has had scarce effect on either teachers or their students (Donaldson, 2009; Toch & Rothman, 2008). However, in 2009 federal education officials sought to expand the role that teacher evaluation plays in school improvement by tying sizeable financial awards in the Race to the Top [RTTT] competition to states' assurances that they would adopt recommended approaches to teacher evaluation. Grissom and Youngs (2015) note the "rapid policy diffusion" that followed (p. 169) as 46 states enacted new teacher evaluation policies (Steinberg & Donaldson, in press), even though only 19 won RTTT grants.

At that time, the problems with traditional teacher evaluation were well-documented. Teachers were unevenly observed and intermittently evaluated, meaningful feedback was in short supply, and few teachers received tenure or were dismissed on the basis of performance (Donaldson, 2009; McLaughlin & Peiffer, 1988; Toch & Rothman, 2008; Weisberg et al., 2009). In response, policymakers, the media, and reformers turned to teacher evaluation as a means to improve schools.

Analysts and advocates who acted as "policy entrepreneurs" (Kingdon, 2002) during the recent enactment of teacher evaluation policies tended to marshal evidence on behalf of their particular definition of the problem to be addressed. Some focused on the need for greater professional accountability, criticizing school officials for failing to hold teachers responsible for their performance or to dismiss those who were ineffective (Thomas, Wingert, Conant, & Register, 2010). These critics urged that evaluation instruments become more discriminating, and that school administrators conduct comprehensive, frequent, and consequential assessments. For example, The New Teacher Project's (TNTP) widely circulated report*, The Widget Effect* (Weisberg at al., 2009), recommended that evaluation instruments include a range of performance ratings so that administrators could identify and dismiss ineffective teachers while rewarding their successful peers.

Other analysts faulted current evaluation systems for not supporting teachers' development and offering them only cursory observations and shallow, irrelevant feedback. Randi Weingarten (no date), president of the American Federation of Teachers argued: "Teacher evaluation in most school districts is not the catalyst for professional growth. It is time for that to change" (p. 2). Other non-profit organizations, including the Aspen Institute (Curtis & Weiner, 2012) and Learning Forward (von Frank, 2013), proposed new systems that would provide teachers with detailed feedback and steady support for improving instruction. Although many reports recommended that teacher evaluation provide greater accountability *and* more support for development, most favored one goal or the other.

Federal RTTT guidelines called for states to address both purposes, by developing "rigorous, transparent, and fair" annual teacher evaluation systems that would "include timely and constructive feedback," "provide teachers with data on student achievement growth," "differentiate effectiveness using multiple ratings," and be used "to inform decisions about staff development, compensation, promotion, tenure, certification, and removal of ineffective teachers" (IES, 2014, p. 5). In crafting RTTT regulations, federal officials relied on two types of policy instruments for "getting the job done" (McDonnell & Elmore, 1987, p. 133), tying *inducements* of federal funding to *mandates* for including specific elements in their evaluation systems.

The federal strategy led to new evaluation policies in most states (National Council on Teacher Quality, 2013), although analysts differ in their accounts of the new laws. Steinberg and Donaldson (in press) reviewed all 46 new state evaluation policies and concluded that most state policymakers took "a developmental stance towards evaluation" by requiring a summative rating to be linked to professional development for teachers, including those judged to be under-performing (p. 13). In contrast, the federal IES (2014) reported that there was convincing evidence of movement toward accountability in the new laws, noting that those most aligned with RTTT priorities "focused on using multiple measures to evaluate teacher performance (30 states); using multiple rating categories to classify teacher performance (31 states) and conducting annual evaluations (25 states)" (p.1).

However, the story of evaluation reform does not end with new state policies. Decades of research about policy implementation show that enacting policy at the federal or state level does not automatically lead to to changes in local practice. (Pressman & Wildavsky, 1973; McLaughlin, 1987; McDonnell & Weatherford, 2013). In the case of teacher evaluation, this suggests that the financial inducements and mandates of RTTT, which exerted such strong influence at the state level, might not hold sway in local districts and schools, especially in states with less centralized education systems. Local officials might disregard the state's requirements, feign endorsement, fail to fund them, or comply minimally.

Even if district officials were to meet the state's requirements, there would be no assurance that school-based administrators would be able or willing to implement them. As Lipsky (2010) explains, it is "street-level bureaucrats," such as principals and teachers, who decide the ultimate fate of policies enacted by higher levels of government. Moreover, because policy entrepreneurs and policy makers at the federal and state levels featured different problem definitions of teacher evaluation during their debate and subsequent drafting of laws and regulations, principals who implemented new policies might favor one or the other problem definition (accountability *or* development) and adopt different practices in response. Ultimately, in order to understand how new teacher evaluation policies affect public education, we must investigate whether and how they are being implemented and experienced by principals and teachers in schools.

With that in mind, we focused this 2014 study on how six schools—including both charter and district schools—located in one Massachusetts city were implementing a 2011 state policy for teacher evaluation. All schools served high-poverty, high-minority communities and had achieved the state's highest accountability rating. These schools were in the second or third year of implementing the policy, based on the state's timetable for roll-out. Through interviews, document analysis, and informal observations, we considered three components of each school's evaluation process—classroom observations, follow-up feedback, and summative ratings of teachers' instruction—examining whether evaluation was used to promote greater accountability, more opportunities for development, or both.

## Theoretical Framework

Recent theory and research about the role of "sense-making" in policy implementation (Coburn, 2001, 2005; Spillane; Reiser & Reimer, 2002) provides an informative perspective for understanding the complex ways in which individual and social cognition shape the ultimate character and impact of a policy as it plays out in schools and classroom. This perspective, as Spillane, Reiser, & Gomez (2006) explain, goes well beyond acknowledging the established truth that "implementing agents" will interpret policies differently (p. 389). Most scholarship about sense-making in policy implementation focuses on curriculum and instruction (Cohen, 1990; Cohen & Hill, 2001; Coburn, 2001, 2005), with the recurrent finding that teachers who implement policies often see much similarity in their current instructional practice and the new policy's requirements, leading them to modestly change their pedagogy, if at all. As Spillane, Reiser, and Gomez (2006) observe, this research "suggests that seeing new ideas as familiar is indeed an obstacle to implementation" (p. 54).

Spillane, Reiser, and Reimer (2002) identify three components of sense-making that together determine what educators do in response to policy directives. The first, "individual cognition," designates the ways in which individual implementers "notice and interpret stimuli and how prior knowledge, beliefs, and experiences influence construction of new understanding" (p. 388). The second, "situated cognition," includes elements of the implementers' social and organizational context that affect their responses. The third, "policy stimuli," refers to messages and signals that implementers receive about policymakers' intentions and how they ought to respond (p. 389).

Sense-making theory has not yet been used systematically to analyze policies designed to change the composition and quality of the teaching force, although it promises to have great utility. A principal might be influenced by all three components of sense-making as she decides how to respond to a new evaluation policy. For example, in past experience, she might have developed skills and beliefs about the demands and potential of evaluation to improve teachers' practice. However, as principal, her work is embedded in the social and organizational context of a school district, which sets standards for how she should implement the policy. Third, she will be aware of the messages conveyed by policymakers in promulgating the law and its regulations. Those “policy stimuli” might focus on the goal of achieving greater accountability or providing more support for development. Given explicit and competing priorities promoted by policy entrepreneurs and policymakers at several levels, a principal might respond to one message, while discounting others.

Further, we must consider teachers’ responses to new evaluation laws, since they, too, bring prior knowledge and beliefs to implementation, are influenced by a particular school and district context, and respond to policy stimuli in the media or communications from their state or district. Teachers might welcome being observed by a knowledgeable, constructive critic (individual cognition), but doubt that their principal has time to complete the required number of observations (situated cognition), and hear conflicting messages about the new policy’s goal (policy stimuli). In response, they might take a wait-and-see attitude about the new evaluation system.

This is complicated terrain for research because interactions occur, not only among the three components that influence individual actors, but also among the actors. However, if we are to inform practice as well as subsequent policymaking, we must acknowledge that complexity and turn our attention to schools and classrooms.

We initially attended to whether the six schools we studied actually implemented the state’s policy—all did. We then adopted an inductive, bottom-up approach to data collection and analysis, where we learned how evaluation was conducted within each school; whether the school’s educators focused on goals of development, accountability, or both; and how administrators and teachers viewed the strengths and weaknesses of their supervision and evaluation process, noting where participants’ views aligned, conflicted, or diverged. In what follows, we first review the relevant research literature. Next, we describe our methods, provide an overview of the Massachusetts policy, introduce our schools, and report our findings. We conclude by discussing the implications of our study and findings for policy, practice, and further research.

## Literature Review

Over a decade ago, value-added analyses by Rivkin, Hanushek, & Kain (2005) and Rockoff (2004) documented wide variation in teachers' effectiveness within schools, suggesting that administrators did not routinely base employment decisions (reappointment, tenure, and dismissal) on evidence about teachers' performance. Researchers, analysts, and the media have attributed this weak role of evaluation to various factors—limitations of the instruments used; irregular and incomplete teacher observations; uninformed summative judgments; and administrators' reluctance to arouse professional or political opposition by initiating teacher dismissals (Donaldson, 2009; McLaughlin & Peiffer, 1988; Toch & Rothman, 2008).

When Weisberg et al. of TNTP (2009) studied personnel practices in 12 districts, they found that in districts with only two rating options (e.g. "effective" and "ineffective") over 99 % of teachers received effective ratings, even though 81 % of administrators and 57 % of teachers said that a teacher in their school was performing poorly, and 43 % said that a tenured teacher in their schools should be dismissed (p. 6). They attributed much of the problem to limitations in the evaluation instruments, themselves.

Proponents of greater accountability have sought to achieve more valid and reliable ratings. The Bill and Melinda Gate's Foundation's 3-year MET Project (2013) found that combining multiple measures—standards-based observations, student surveys, and student achievement gains—yields the best estimates of teacher effectiveness, as measured by test scores. However, research has just begun on how evaluations based on multiple measures—especially those including student achievement—are being implemented.

Prior to recent enactment of state teacher evaluation policies, researchers studied new approaches to evaluation in a selection of local districts and charter management organizations (CMOs). In Cincinnati, Taylor and Tyler (2012) examined the effects on individual teachers' instruction of being evaluated by trained and experienced evaluators—including administrators and peers—who combined "multiple, detailed classroom observations and a review of work products" with face-to-face feedback (p. 2). Mid-career math teachers' effectiveness improved and those gains persisted, and even increased, several years after the evaluation cycle ended. The authors suggested this was due to the rich feedback and encouragement to improve that teachers received.

Donaldson and Papay (2015) interviewed 95 administrators and teachers in New Haven, CT, during the second year of implementing TEVAL, a system in which evaluators assessed teachers based on classroom observations and students' academic growth. TEVAL offered support for ineffective teachers' development and, if that failed, a clear pathway to dismissal. Prior to TEVAL, no tenured teachers had been dismissed for poor performance. Two years after its enactment, all teachers who were notified that they would be dismissed resigned (1% of tenured teachers, 3% of non-tenured teachers). Overall, teachers said that TEVAL focused their attention on test results, but did not lead them to change their instruction.

When Donaldson & Peske (2010) interviewed teachers about evaluation in five charter schools of three CMOs, they heard very positive comments about weekly or biweekly observations, followed by coaching sessions with their evaluator. Evaluation served primarily as a process for professional growth, with far less emphasis on the summative assessment. Similarly, in 15 New York City charter schools studied by Dobbie and Fryer (2011), students of teachers who received formal and informal feedback ten or more times per semester showed higher learning gains than those of other teachers.

In several studies, scholars explored factors that influence both principals' and teachers' responses to an evaluation process using standards-based instruments. Notably, they found that both the principals' approaches and the teachers' responses varied widely across schools. For example, Kimball and Milanowski (2009) interviewed two sub-groups of principals, whose ratings correlated either well or poorly with their students' achievement data, and considered whether differences in principals' motivation, knowledge, skill and/or school context explained differences in the validity of their ratings. However, no clear explanation emerged from their data. Sartain and colleagues (2011) interviewed 37 Chicago principals trained to conduct observations using a standards-based framework and concluded that the improved evaluation tools and training could support principals in conducting reliable assessments and engaging teachers in reflective, developmental conversations. However, teachers suggested that training was not enough and that success depended on their principal having not only strong knowledge of the framework, but also well-developed coaching skills, and high engagement in the process. Similarly, O'Pry and Schumacher (2012) interviewed new Houston teachers and found that their perceptions about evaluation were determined largely by the value they thought their principal placed in the process.

In our earlier study of six low-income schools in one large, urban district (Reinhorn

& Johnson, 2014), we also found wide variation in principals' approaches and teachers' responses to teacher evaluation, even though the principals used a common standards-based instrument and encountered the same "policy stimuli," through memoranda and training from the central office. One principal concentrated on building dismissal cases for a few weak teachers. Four others treated evaluation largely as an inconsequential, bureaucratic obligation. In only one school, did the principal integrate formative and summative assessment in a process that teachers said helped them to improve their teaching. Subsequently, Kraft and Gilmour (2015) interviewed 24 principals in one large urban district about their evaluation practices. Principals said that, in discussing feedback with teachers, they focused largely on the evidence that supported their ratings, but had little to suggest about what teachers should do to improve.

Throughout these studies, we find evidence about the crucial role that principals' knowledge and beliefs play in how they use evaluations. We also learn about the role of district context as principals decide where to focus their efforts and whether to commit time to achieving the policy's advertised purposes. For their part, teachers also were influenced by various elements of sense-making as they responded to evaluation, including what they believed their principal's goals were, how well they thought their principal understood instruction, and whether their principal could offer constructive feedback.

### Early Evidence about the Implementation of New Evaluation Policies

State evaluation policies adopted after RTTT are still at early stages of implementation, but early research suggests similar responses across states and large districts. These studies find that overall principals tend to focus on development rather than accountability and that contextual conditions, especially constraints on administrators' time, limit what they can do.

Donaldson and Casey (2015) studied the pilot implementation of Connecticut's new evaluation policy and found that principals had trouble finding time to complete the six required observations (three formal, two informal) of all teachers. Still, teachers appreciated receiving more feedback than they had experienced in the past. In a similar study New Jersey principals also reported that their policy's requirements for observations exceeded what they could complete (Firestone, Nordin, Shcherbakov, Kirova, & Blitz, 2014). Whereas Connecticut teachers expressed trust in their system, many New Jersey teachers voiced distrust and concern about job security. Only about one third of New Jersey teachers said that the process "helped them improve some aspect of their teaching" (p. 22).

Drake et al. (2015) interviewed central office personnel in five large, urban school systems located in different states and learned from district administrators that principals focused first on developing teachers throughout the school year and only subsequently focused on dismissing those who did not improve. Dismissal, they found, was a "byproduct of a long support process" (p. 119). However, principals also found that the time required to complete the dismissal process was a barrier to pursuing it.

Since RTTT, the dominant policy stimuli have encouraged using evaluation for accountability, although evidence suggests that principals continue to rate few teachers unsatisfactory (Taylor, 2015; Doherty & Jacobs, 2015). McGuinn (2012, 2015), who has tracked implementation of new evaluation policy by SEAs in six states, contends that it "is necessary to move the conversation from the punitive—focusing on getting rid of a small number of bad teachers—to the productive—using better information to improve the instruction of all teachers and creating a continuous improvement model" (p. 14).

Across these and earlier studies, researchers document considerable school-to-school variation in evaluators' practices and teachers' responses. We expected to find comparable variation across our sample of schools, but instead found surprising similarity in these administrators' approaches, all of which featured development over accountability.

## Methods

This article is based on a qualitative, comparative analysis (Maxwell, 1996) of data drawn from a larger study examining how six high-performing, high-poverty, urban schools attract, develop, and retain teachers. Here we focus on teachers' and administrators' approaches to and experiences with teacher evaluation in their school. We ask:

1) How do principals and other administrators in these six schools interpret their state's teacher evaluation policy and how do they implement it? Do their priorities and practices differ across schools? If so, how?
2) How do teachers describe and assess their experience with evaluation? Do their descriptions and judgments about the process vary within their school or from school to school? If so, how?

### Sample of Schools

Our sample includes six elementary and middle schools (three charter and three district) all located in Walker City, Massachusetts. (Pseudonyms are used for all places and individuals.) All schools served high-poverty populations (70% or more eligible for free or reduced-price lunch), with high proportions of students of color. Although the state had previously intervened in three schools because of chronically poor performance, all had achieved the highest rating in the state's accountability system by 2013-2014 and were widely viewed as high-performing. All six principals and CMO heads whom we asked to participate in the study agreed. (For descriptive statistics of the sample schools see Table 1.) The purposive nature of our sample allows us to examine the practices of this set of schools and to consider the implications of our findings for others, but does not permit causal inferences or generalizations beyond the sample.

[Insert Table 1, *Selected Characteristics of Six Sample Schools*, here.]

## Data Collection

In order to understand how principals and teachers understood and implemented the evaluation policy, we conducted interviews, analyzed documents, and informally observed in the schools. Between March and June 2014, we conducted 142 semi-structured interviews with teachers, administrators and other staff, such as curriculum coaches and program coordinators. We solicited teachers' participation by email and flyers, and followed up on recommendations from those we interviewed about others we should contact. We interviewed between 33% and 56% of the teachers at each school, depending on its size and complexity. (For descriptive statistics about the interviewees, see Appendix A). We used semi-structured protocols to guide our interviews, ensuring that data would be comparable across sites and across interviewers. We promised participants confidentiality and anonymity. All interviews were recorded and transcribed. In the course of visiting the schools to conduct interviews, we informally observed practices in classrooms, corridors, and offices, which we recorded in field notes. We also gathered and analyzed relevant documents, including teacher evaluation frameworks and rubrics; teacher handbooks; school, district, and state policies; and examples of observation feedback to teachers.

## Data Analysis and Validity

We wrote a structured thematic summary after each interview and then analyzed sets of thematic summaries to identify common themes and differences within and across sites. In developing thematic codes, we supplemented the *etic* codes drawn from the literature (e.g., "adminteach" for quotes referring to the relationship between administrators and teachers), with *emic* codes (e.g., "demands" for quotes about teachers' professional responsibilities within the school) that emerged from the data (Miles & Huberman, 1994). After achieving inter-rater reliability by simultaneously coding a subset of transcripts and then calibrating our interpretation of the codes, we coded each transcribed interview using the software, Dedoose.

Based on interview data and document analysis, we created analytic matrices (Miles & Huberman, 1994) so that we could systematically consider our research questions about practices within and across schools. We addressed risks to validity by returning to the data to review our coding and emerging findings, and seeking rival explanations or disconfirming data (Miles & Huberman, 1994). We also conducted member checks by sharing initial findings with principals from all schools and by providing all participants with links to our working papers. In each case, we invited participants' responses.

## The Massachusetts Evaluation Policy

The teacher evaluation policy was developed in 2011 by a 41-member Task Force, which included public education administrators, teachers, and representatives from universities, foundations, business, unions, and non-profit agencies. Their report explains that they sought to "transform educator evaluation from an inconsistently applied compliance mechanism into a statewide catalyst for educator development and continuous professional growth" (p.5). The policy's 5-step cycle for continuous improvement called for teachers to set “specific, actionable and measurable” goals for improving their practice and students’ learning. Subsequent state regulations, which gave evaluators the final say in what teachers’ goals would be, required them to conduct mid-year formative and end-of-year summative assessments, which included ratings on the four standards of professional practice, an assessment of progress toward meeting their goals, and an overall rating.

In his memorandum recommending the report to the State Board (Chester, 2011), the Commissioner of Education urged that in establishing regulations, Board members balance the proposed policy’s focus on teachers’ development with the schools’ obligation to ensure accountability—to “dismiss educators who, despite the opportunity [to improve], continue weak performance.” Thus, the policy stimuli signaled the importance of both development and accountability. Following the policy’s enactment, state officials conducted an ambitious program of information and support for districts, including a model evaluation system, contract language, and training modules.

## Three District Schools and Three Charter Schools

*Dickinson Elementary,* a century–old district neighborhood school served a largely immigrant student population. Well regarded within WCSD, Dickinson experienced very low teacher turnover; in 2014, over half of Dickinson’s teachers had taught there more than 20 years. Dickinson’s Principal Davila complied with the WCSD teachers contract, as well as other district and state policies. She had no special autonomy in staffing or scheduling.

*Hurston PK-8 School* and *Fitzgerald Elementary School PK-5*, also part of WCSD, each had been placed in turnaround by the state in 2010 because of persistent failure. Under RTTT guidelines, the newly appointed principals, Hurston’s Roger Hinds and Fitzgerald’s Sharon Forte, had the right to replace all teachers, but could retain no more than half. Hinds replaced about 80 % of the staff and Forte replaced about 65 %. Each school continued to enroll students from the same local community as before turnaround. Subsequently, both showed substantial growth on the Massachusetts Comprehensive Assessment System (MCAS), allowing them to exit turnaround status at Level 1 of the state's accountability rankings.

After turnaround, both Hurston and Fitzgerald remained WCSD district schools, although each retained significant school-based control of its organization and management, making it possible to continue many of its initiatives. Prior to turnaround, Hurston PK-8 had gained special status in WCSD, allowing the school to hire and transfer out staff, choose its curriculum, allocate its budget and set its schedule. As Fitzgerald PK-5 emerged from turnaround, Principal Forte successfully applied to become a state Innovation School within the district, which brought with it many of the management autonomies previously available during turnaround.

*Naylor Charter School (K-8)* and *Rodriquez Charter School* (PK-8) opened in Walker City 10 and 20 years earlier as freestanding state charter schools. In 2014 Naylor was one of three schools in the expanding Naylor Charter Network. Although located within WCSD boundaries, these schools were exempt from local district policies.

*Kincaid Charter School* (6-8) was part of The Kincaid Charter Network, a CMO the state selected to restart a failing WCSD middle school in 2011. All of the school's teachers could reapply for positions in the new charter school, but few did and none was rehired. All administrators, teachers, and staff were new to Kincaid Charter when it opened, although approximately 80% of the students returned, a higher proportion than typically re-enrolled in prior years. As a restart school, Kincaid functioned as an in-district charter school; the local union represented Kincaid's teachers, whose pay aligned with WCSD's negotiated scale. However, the school was exempt from other contract provisions. Within two years, Kincaid Charter made significant gains in student test scores and achieved a Level 1 rating from the state.

## Findings

Across the six schools, administrators interpreted the state evaluation policy with remarkable similarity, all giving priority to the goal of development over accountability. Teachers concurred that evaluation in their school was meant to improve their performance, which they strongly endorsed. All six schools not only complied with the new regulations of the evaluation law, but went beyond, by providing teachers with frequent observations, feedback, and support. However, as a result of being a traditional, charter, turnaround or restart school, each principal responded to a distinct policy context, which led to variations in how evaluation worked. Interviews revealed how the various components of sense-making—individual cognition, situational cognition and policy stimuli—combined to influence school-based implementation of the state's new policy.

Districts and schools could "adopt or adapt" the state's comprehensive Model System or "revise" their existing evaluation system to meet the new regulations. WCSD, and therefore the three district schools, as well as Rodriguez Charter had adopted the Model System. Kincaid Charter and Naylor Charter revised their existing process to meet the state's new regulations. All schools used detailed, standards-based frameworks for observations and assessments. Teachers participated actively in setting individual goals for student performance and professional practice, and in completing self-assessments prior to formal evaluation meetings. Every teacher eventually received a summative rating at one of four levels of proficiency. At the time of this study, evaluators were not yet required to use student achievement data in the evaluation process.

The new evaluation policy was not the only policy that influenced how principals implemented evaluation. After the state introduced MCAS in 1993, it took an increasingly active role in monitoring school performance. In 2012, following RTTT, it began to rate schools and districts from Level 1 (highest-performing) to Level 5 (lowest-performing). The Commissioner could designate chronically low-performing schools at level 3 or 4 for turnaround, restart, or closure. If a turnaround school failed to improve on its required timeline, it could be placed at level 5, which triggered state receivership.

Most schools worked hard to avoid public censure for receiving a low rating or, worse, being designated for turnaround or restart. However, if the state intervened, the process of re-opening a school with a new principal and newly constituted faculty provided flexibility and management opportunities that other schools did not have. Principals not only had the right to hire, fire, and transfer their teachers, they also had discretion in allocating teachers' time. Further, the school could apply for additional resources through federal School Improvement Grants, which these schools often used to extend the school day, employ additional staff (including administrators), and fund more professional development time for teachers. State-sponsored charter schools already had authority to hire and fire staff and allocate their time, much as principals of turnaround and restart schools could. Also, charter schools could extend the school day and year as well as raise additional funds to support their program and operations.

As the principals in our study implemented the state's new evaluation policy, they did so with attention to the broader context of laws that imposed accountability and threatened sanctions for failure, while also providing managerial flexibility and additional resources for some schools under the state's control. Principals were further influenced by their own beliefs and knowledge about teacher evaluation and school improvement (individual cognition), resources and training within their district or CMO (situated cognition), and by messages generated in the national debate about evaluation, federal RTTT guidelines, and the MA law, guidelines, and support for implementation (policy stimuli).

**The Schools Focused Primarily on Development**

In all six schools, administrators viewed the primary purpose of evaluation to be developing their teachers, many of whom they had hired. Samantha Nelson, Naylor Charter's network head, explained that a commitment to improvement called for frequent observations: "We do believe that our whole mission is to be a human capital organization. We are here to develop our kids. We are here to develop our teachers. We are here to develop our administrators. This is what we do and what we're all about." As a result, she said, administrators focused their time conducting observations and providing feedback: "[W]e think that the most transformational thing is just being in people's classrooms, talking with them afterwards."

Across schools, when we asked about teacher evaluation, administrators began by describing their approach to formative, rather than summative, evaluation. An administrator at Kincaid Charter explained, "We believe that teachers, or just people in general, grow with immediate feedback and real-time instruction on how they are performing and giv[ing] them an opportunity to fix that in the moment." Kincaid's principal Daniel Kain realized that some teachers would encounter more challenges than others, but he was confident that those in his school had the expertise needed to support them: "If we have teachers who are struggling, it's often times … rooted in a lack of skill. Our job as coaches is to help them with that." Administrators in this sample provided detailed explanations about how, as one Fitzgerald evaluator said, they "coach [teachers] or find them the help they need."

Teachers confirmed their administrators' accounts, explaining that evaluation focused on promoting their growth. A Naylor Charter teacher said she appreciated administrators at her school "continuing to develop [her] as a professional." Her colleague said that teachers wanted to improve: "[I]n order to be an employee here, regardless of if you're an academic teacher, co-curricular teacher, even a staff member, you need to want feedback … to get better… [to] help [our] kids."

Teachers in all six schools said that the evaluation process was embedded in a professional culture that promoted continuous improvement. Many described their school much as this Rodriquez instructional coach did: "[T]here is a culture here that is about continually getting better…that means that every teacher, whether they're getting feedback from an administrator or not, is trying to get better in their own practice."

**Observation and Feedback Practices that Supported Development.** Under the state's Model System, adopted by WCSD and Rodriguez Charter, the number of required observations depended on the teacher's summative rating in prior years. Principals had to observe a new teacher or a teacher who had been rated "unsatisfactory" one time in an announced visit and four times in unannounced visits. A returning teacher with a history of "proficient" or "exemplary" ratings had to be observed only once, unannounced. A teacher who had received a "needs improvement" rating had to be observed at least twice, unannounced.

In this sample, most teachers described an intense cycle of observations, followed soon after by written or oral critique and recommendations from their evaluator. Approximately, 40% of the 99 teachers we interviewed said they were observed and received feedback at least twice per month. Approximately 20% estimated that they were observed and given feedback between 5 and 10 times per year. The final 40% estimated that they had been observed 1 to 4 times per year, consistent with state and district policies. Although all schools met or exceeded the state requirements, the frequency of observations varied within schools, with novice teachers and new hires being observed more often than others.

Kincaid Charter and Naylor Charter administrators expected that every teacher would be observed and provided face-to-face feedback at least twice per month and all teachers interviewed said that evaluators met, and sometimes exceeded, that standard. At Hurston K-8 and Rodriguez Charter, administrators aspired to observe every teacher and provide feedback at least once per month, although participants said their school did not have the resources to maintain that level of supervision for all teachers. However, all administrators routinely conducted “walk-throughs” for quick observations, often providing feedback. Dickinson and Fitzgerald teachers said their principals spent a great deal of time in classrooms throughout the school, but most described receiving formal feedback no more than a few times per year, which was consistent with the district’s requirements.

It is notable that principals in the six schools expressed similar beliefs about the benefits of frequent observation and feedback, which was not the case in studies discussed earlier, where principals' views varied widely. Moreover, all were recognized for being strong, experienced teachers, themselves, and therefore they brought to the process not only beliefs about the benefits of developing teachers, but also knowledge and skills about how to do so.

Three schools (Naylor Charter, Kincaid Charter and Hurston K-8) had sufficient administrative resources available so that principals could spend their time observing teachers and providing feedback, while other administrators in their school handled responsibilities such as student discipline, interaction with families, and operations (e.g., building maintenance, bus schedules, data analysis, and budgeting). The Director of Operations at Hurston K-8 explained, "My role has been to block and tackle so that [the evaluators] can spend their time in the classroom coaching teachers and at [teacher] team meetings." Notably, however, Hurston's administrative team was the same size as Kincaid's, although it served twice as many students and teachers. Therefore, regardless of Principal Hinds' intentions, Hurston's evaluators could not provide the same level of intense supervision for all teachers as their counterparts could at Kincaid and Naylor charter schools. Having at most a principal and assistant principal—and in the case of Dickinson only a principal—they could not hand off management responsibilities to others and spend their time in classrooms. Therefore, in addition to these principals' knowledge and beliefs about instruction and evaluations (individual cognition), the realities of their context (situated cognition) also affected what they could do. Whatever their beliefs and intentions, principals with less administrative support coped with more demands and greater constraints on their time.

**Teachers' Responses Were Overwhelmingly Positive.** Although some principals expressed concern about not being able to observe teachers often enough, most teachers endorsed the observations and feedback as a positive part of their professional experience. They expressed their appreciation with phrases such as "hugely helpful" or "super supported." One Naylor teacher, in her seventh year of teaching and her fifth at the school described "the constant feedback" she received as a highlight of her job: "I constantly feel like I'm getting better." Similarly, a Fitzgerald teacher said that evaluation kept her "on [her] toes" and "helped [her] to do better as a teacher." A Dickinson teacher echoed, "It's helpful always. A second person can notice things that you, yourself, in the job miss." Others suggested that administrators demonstrated their investment in developing teachers by often observing their classes. A third-year teacher at Rodriguez Charter said, "Just the fact that my administrators are in my classroom on a weekly or bi-weekly basis, I think shows a lot. It means that they care, and they're here to help us." Across various levels of experience, teachers said that evaluators had greater credibility and gained a better understanding of individuals' professional experience and struggles if they observed them teaching often. A teacher at Rodriguez, with 12 years of experience, explained, "He knows my flaws. He knows what I need to work on. He knows me better than I know myself as a teacher."

Many teachers described a professional culture within their school that encouraged them to view evaluation as a developmental process. One from Naylor Charter explained:

> [I]n my old school …you'd find out they were coming in [to observe]. It was like you were ready for a performance. You had to do it perfectly and then they never came in again until three or four months later. [Here], they're just always in and out of the room, so it's nice. It's a good way to just always keep getting better.

A colleague offered a similar perspective: “When I know something isn’t going well, I will ask to be observed so that I can get help on that. That’s totally the mentality here. I don’t like someone seeing me doing something wrong.” However, she noted that she would “prefer that … [to] not getting any guidance on it.” Therefore, teachers’ responsiveness to the evaluation process was nurtured by their beliefs about how they could improve (individual cognition), as well as evidence from their school context about the instructional expertise of their evaluators and a strong professional culture that reinforced the value of investing in their development (situated cognition).

**Teachers reported receiving detailed, helpful feedback.** Most of these teachers said that their evaluator provided detailed feedback about a range of topics including classroom management and pedagogical strategies. A teacher at Naylor Charter said that her supervisor had helped her improve the questions she asked during read-alouds so that she could promote higher-order thinking among students. Naylor administrators provided written feedback on a Google Doc shared with the teacher, which teachers appreciated because it helped them to see their progress over time. A Kincaid Charter teacher described the feedback she received about the ratio of teacher talk to student talk during her lesson and a Fitzgerald teacher said that she received helpful feedback about pacing lessons. Hurston K-8 evaluators emailed their feedback within 24 hours and recorded observations on a Google Doc shared with the administrative team. Teachers repeatedly described the post-observation feedback as timely, specific and relevant. Several showed us the comments they received. Principal Hinds wrote to one teacher about the pacing of a lesson and to another about the interaction during class discussion as the teacher responded to every student’s contribution before the next student spoke. In both cases, he offered suggestions that the teachers found helpful.

Many teachers said that the observation and feedback process had led them to change their pedagogy and that their practice was improving. A Kincaid Charter teacher with six years experience said that she had become "a drastically better teacher" in the three years that she had worked at the school, "because it's been this really close cycle of being observed and then feedback on what to work on, and then observed again and then feedback again." An early elementary teacher at Rodriguez Charter, who had ten years of experience, described how his principal provided him with observational feedback over time, which supported him in dramatically shifting his instructional approach.

> She kind of said, "Why don't you think about doing this, that and the other thing?" I said, "Okay" and that first two, three, four weeks of changing my entire teaching style was a disaster. …I started tweaking it and figuring it out and she would come in and observe and critique and give good positive comments and negative ones. …Looking back I can't even imagine how much of a disservice I was doing to kids back then in the way that I was teaching.

### Integration With Other Practices

Each of these principals had an integrated strategy for improving teachers' practice across the school. Evaluation did not stand alone, but rather, was coordinated with other professional learning opportunities, such as instructional coaching, teacher teams, whole school professional development, and peer observation. Teachers also reported that some administrators remained informed about their professional practice by reviewing unit and lesson plans and participating in team meetings, which focused on data analysis and curriculum planning. In explaining the support they received, teachers often did not distinguish between practices that were part of the evaluation system and others; they considered them all as part of an ongoing, integrated improvement process. However, many identified classroom observations and feedback as the most valuable component of the process. They expressed confidence that administrators and coaches would provide support as teachers responded to their feedback (situated cognition). For their part, principals believed that it was their responsibility to support teachers' development (individual cognition) and they organized their time to make that happen as best they could.

The goals that teachers were required to set in the evaluation process served to integrate elements of the individual, team and whole school improvement processes. The state's Model System stated: "Connecting individual educator goals to larger school and district priorities is critical to effective implementation. Strong vertical alignment between individual, team, school and district goals will accelerate progress on the goals. "In the four schools using the Model System (Dickinson, Fitzgerald, Hurston K-8 and Rodriguez Charter) teachers explained that their goals were intentionally connected to school-wide and team-based goals. An administrator at Hurston K-8 explained the advantage of explicitly linking these processes. "So there's… an alignment from the individual to the team to the school that makes sense to people, and it doesn't feel like they're pulling [evaluation] goals out of the hat." Thus, as administrators in these schools implemented the evaluation policy, they were influenced not only by beliefs and context, but also by policy stimuli, conveyed by the state through its Model System.

### Evaluation For Accountability Was Grounded in Evaluation For Development

The school leaders' focus on teachers' development was not seen to be in tension with the summative evaluation process, which included mid- and end-of-year meetings to discuss teachers' ratings on the evaluation rubric and their progress toward reaching designated goals. Participants realized that formal evaluation could be used to inform current and future employment decisions, but that did not dominate the process.

Teachers widely said that the formal evaluation process provided an accurate assessment of their professional practice. Unlike ongoing formative supervision, which usually focused on no more than a few issues at a time, summative evaluation was comprehensive and detailed. Evaluators rated teachers' performance on all four standards, each including a number of indicators defined by specific elements and descriptors. In five schools, accompanying rubrics depicted typical performance at each of four level of accomplishment. Some teachers said that they respected the fact that even the summative rating process also encouraged improvement. This was especially true when teachers believed that their evaluators had a deep understanding of learning and teaching. A Naylor teacher explained that she was graded on "a rubric from 1-4, just like the students are." She noted that, despite receiving "mostly 1.5s, some 2s and a 3," she was not discouraged, although "in another context, I would have felt like they were starting a paper trail to fire me." She explained that her current administrators had different expectations for beginning teachers. "They expect their first-year, maybe even second-year, teachers to be working hard, but not really mastering all the things they want you to master." Overall, teachers trusted that the summative process, like the formative process, was intended to support their growth and therefore they could accept tough assessments of their practice.

**"No surprises" in formal evaluations.** Many teachers in the sample described formal evaluation processes as an outgrowth of day-to-day supervisory practice. One teacher described formal evaluation as “just a tiny piece of what we already do on a daily basis.” Another teacher echoed many others in explaining that the summative evaluation process “shouldn’t be a big deal. It really hasn’t [been].” Another expanded: “I know exactly what my goals are and what I’m doing, so it wasn’t surprising how she graded me. I graded myself really hard, but I knew what I was working on, so it made sense to me.” This teacher’s individual beliefs about the legitimacy of the rating system led her to take it seriously, although the situational component of sense-making also came into play; she could be confident that acknowledging her shortcomings would not lead to reprimand. As Principal Hinds explained, the administrators’ intentions matched the teachers’:

> I think evaluation without ongoing supervision is meaningless. It becomes only the way that you terminate employment. And so my belief is that I and every member of my administrative team needs to be in classrooms all the time, giving feedback, asking questions, pushing people. And then all of that just gets rolled into an evaluation. No surprises.

Nonetheless, across the sample, teachers and administrators said that evaluation could be used to hold teachers accountable for meeting professional expectations. Teachers believed that, when warranted, evaluators did give teachers low ratings on summative evaluations, which could lead to dismissal. At several schools, teachers and administrators spoke of teachers who were on official improvement plans with goals that they had to meet in order to keep their position. Administrators also told of teachers who were not offered a position the following year or, in charter schools, had been dismissed mid-year. This contributed to a sense of accountability and made the evaluation process a serious one, but it did not seem to generate fear or undermine the teachers’ trust in their evaluator or the system. Therefore, the fact that these administrators were intent on developing their teachers did not mean that they avoided dismissing or counseling out those they thought should leave.

For teachers in most schools, summative evaluation grew out of frequent informal and formal observations and feedback and, therefore, they granted it legitimacy, which they might have withheld if classroom visits were rare or they found feedback vague or off the mark. Based on their acquired understanding of their school context, teachers did not expect or want a rubber-stamp of approval; nor did they think that falling short of the highest rating would be the first step out the door. Teachers widely expressed confidence that they were beneficiaries rather than casualties of their school's evaluation process. By contrast, teachers surveyed or interviewed in the studies discussed earlier differed in whether they thought evaluation in their school served either them or their students.

## Shortcomings in Supervision and Evaluation Processes

Despite the overwhelmingly positive views of supervision and evaluation, teachers and administrators encountered challenges to implementing the policy. Among the most important were mismatches by subject between evaluator and teacher, and the demands that frequent observations placed on evaluators' scarce time. These situational elements influenced teachers' beliefs about the quality of their evaluators' feedback and assessment. Both administrators and teachers thought that their evaluation process could be substantially improved by successfully addressing those limitations.

**Mismatches between evaluators and teachers.** Across schools, teachers expressed confidence in their evaluator's knowledge about classroom management and general pedagogy. However, even teachers who respected their evaluators' pedagogy, sometimes expressed disappointment that they lacked experience teaching their subject and, therefore, could not make sound subject-specific recommendations. For example, a middle school math teacher at Hurston K-8 said that, although her administrator's comments were "affirming," she found her math colleague's feedback to be more helpful. "He just knows more about the content. He can tell if students are understanding or not a little bit more than [administrators] can because not everybody's an expert in everything." A history teacher at Kincaid Charter, who was supervised by a former English teacher, said that his feedback often focused on how to teach writing through history, but neglected the "nitty gritty of history." At Fitzgerald Elementary, Rodriquez Charter and Kincaid Charter, teachers of students with special needs expressed concern that their supervisors did not have experience or knowledge specific to special education. One, whose students had learning disabilities, called the feedback "very standard…cookie-cutter." Another explained, "I think there's still huge amounts of growth I could make, but it's hard accessing that growth when the people up ahead of you don't know what you're doing."

Several administrators acknowledged that they could not provide pedagogical advice in every subject, at every grade level. At Fitzgerald Elementary and Rodriguez Charter, instructional coaches supported teachers in planning and teaching mathematics and literacy, but those content experts did not conduct formal evaluations.

**Insufficient time**. Second, participants said that evaluators lacked the time they needed to provide comprehensive evaluation for all teachers, confirming findings in studies discussed earlier. Virtually all teachers we interviewed were grateful to receive observations with feedback, but some wanted more than their supervisors could provide. As one Fitzgerald teacher said, "I think it would be a lot more powerful if administrators were able to be in the classrooms a lot more." Principals at Dickinson, Fitzgerald, Hurston, and Rodriguez talked about the daunting demands of conducting frequent observations and providing detailed feedback for all teachers. Most evaluators had between 15 and 20 teachers to supervise, but some had more. Principal Hinds at Hurston had the most—39. Principal Forte at Fitzgerald Elementary said, "We just can't keep up. We're lucky to have two of us [principal and assistant principal]." Similarly, an administrator at Rodriguez Charter said, "I have 20 people I evaluate and supervise, and it feels like too many to me. I'm always thinking, 'Oh, I haven't been there for so long!'"

At these schools, teachers generally believed that administrators spent more time supervising new and struggling teachers than proficient, experienced teachers. Veterans understood why novices' learning needs took precedence, but nevertheless said they wanted more support, since they knew they could improve. For example, a Hurston teacher with ten years of experience said: "I would like more feedback, [from]someone who knows my classroom, has seen Student A in October and now can tell me how Student A progressed in March." Her colleague with nine years of experience wanted to be observed more often so that she could have in-depth discussions about her "delivery of instruction," such as, "Did it make sense to do that activity … in groups?" Although these teachers pointed to limitations in the current evaluation process, they still appreciated their school's focus on development and valued the feedback they received.

## Discussion

Much recent policy research focuses on reporting outcomes, rather than illuminating the process of implementation. However, those who make policy and those who implement it need to understand not only whether a policy has its intended effects, but also how those effects are achieved, and what factors promote or compromise successful implementation.

Drawing upon sense-making theory (Spillane; Reiser & Reimer, 2002) and employing qualitative methods, we conducted a comparative case analysis in order to learn how one state's teacher evaluation policy was implemented day to day. By choosing to study six successful schools that serve students from high-poverty communities, we hoped to identify effective practices that others might learn from and use. The Massachusetts teacher evaluation policy, like those in many states, specified two purposes, developing teachers' professional skills and increasing accountability in employment decisions. However, the policy gave priority to the goal of development, which the state further reinforced with its model program and numerous other supports for implementation.

Informed by sense-making theory, we found that as they implemented evaluation policy these six principals brought both knowledge and skills about good teaching and a commitment to use strategies that would support teachers' development (individual cognition). They also had a clear understanding of what their school's particular policy context encouraged and allowed (situated cognition). Further, they responded to the state's policy stimuli, which focused on teachers' development. This focus, conveyed in the report of its Task Force, its Model System, and additional training and supports, aligned with these principals' professional priorities. Although we saw no evidence that the policy stimuli dictated their focus on development, we did find that they were influenced by some of the state's supports, especially its Model System, which four schools had adopted. To the extent that the principals raised concerns about how to provide an evaluation system centered on teachers' development, it was because of the time required to do that well, especially given several schools' limited administrative capacity.

Across all schools, teachers affirmed their principal's commitment to developing teachers and reported that they received frequent, useful feedback about their instruction, which they said helped them to improve. They recognized that poor performance and failure to improve might lead to dismissal, but they expressed confidence in the validity and fairness of the summative assessments they received, largely because those assessments were grounded in frequent observations by evaluators with deep knowledge of instruction, detailed feedback, and professional support from various sources, including instructional coaches. Further, by setting goals and completing the self-assessments that were called for by the state's model program, teachers played an active role in the process leading to their annual formal assessment. As they explained their school's evaluation policy, teachers widely expressed confidence in the knowledge, skills, and good intentions of their evaluator (individual cognition), suggested that their school's developmental approach to evaluation was well-intentioned and useful (situated cognition), and judged the state's approach, in which they willingly and actively participated, as legitimate (policy stimuli). Together, principals' and teachers' beliefs and knowledge, their understanding of the context in which they worked, and principals' professional commitment to developing teachers' talent contributed to effective implementation of the evaluation policy.

Importantly, this sample of successful schools is unique, and it would be foolhardy to imply that a similar teacher evaluation policy could or would succeed in any or all contexts. Various factors combined to determine the fate and effects of the Massachusetts policy in these schools. Some emanated from the state, including the fact that the policy was informed by a broad array of interests and a rich set of skills among Task Force members and their advisors. Schools benefited from the state's model program and additional supports. Other factors embedded in both state and local policy influenced implementation, such as the fact that at one time or another, most of these principals had the flexibility to choose their teachers. They invested substantially in an intensive, informative hiring process in order to ensure that teachers were sufficiently skilled and eager to improve. Other policies, including those that authorized charter schools and those that empowered the state to take over failing schools for turnaround, gave several of these principals control over key elements of teachers' work, such as the length of the school day and allocation of teachers' time. Also, schools in turnaround and restart were eligible for supplementary grants, which they used to fund additional administrators or more time for teachers' professional development. Unfortunately, some of these autonomies and benefits—especially supplementary funding—disappeared when turnaround schools exited state control, making it difficult to continue funding programs that many thought were worthwhile. Although most other principals do not have the same degree of flexibility or access to comparable resources, this study allows us to see how different allocations of autonomy and funding granted by one set of policies contribute to successful implementation of another policy.

## Implications for Policy, Practice, and Research

Readers will see in our findings many implications for how evaluation policy can be developed and successfully implemented in a range of settings. Some lessons have broad relevance, while others depend on specific features of districts and schools.

**Policy.** First, this study provides strong evidence that an evaluation policy focusing on teachers' development can be effectively implemented in ways that serve the interests of schools, students, and teachers. Also, these cases suggest that the goals of development and accountability are compatible when summative evaluations are well grounded in the observations, feedback, and support of a formative evaluation process.

This study also reveals the important role that state education officials play in setting the direction of implementation both before and after a policy is enacted. In this case, the state relied on capacity-building rather than mandates to promote effective implementation (McDonnell & Elmore, 1987). These schools clearly benefited from the Model System and other supports that the state provided, which increased principals' opportunities for agency and leadership as they used evaluation to improve their school. It is also notable that, in implementing the state's evaluation policy, five of these schools benefited from the spillover effects of additional policies, including charter school laws and school accountability regulations, which expanded principals' autonomy over staffing and the resources they had to build administrative capacity. States might be more deliberate in mapping the relationships among current policies, making their benefits more widely available, especially to schools with extensive needs.

**Practice.** These case studies of successful schools demonstrate the pivotal role of the principal in implementing evaluation policies. Unfortunately, districts often do not assign their best principals to the schools that need them most. Our findings suggest that doing so is probably the most important thing district officials can do to ensure that teacher evaluation will be a constructive, productive process.

A principal who knows instruction well can engage the teachers in a process of inquiry, self-reflection, and improvement that benefits the entire school. However, principals must help teachers see opportunity in a comprehensive evaluation system and feel confident about seeking help, taking risks, and acting on good advice. Further, they can amplify the benefits of evaluation by integrating it with other components of their professional growth system (e.g. instructional coaching or teacher teams) and thus provide ongoing, comprehensive support for instruction. We concluded that, in addition to knowing instruction well, principals must recognize that, among their many important responsibilities, selecting teachers is probably the most consequential. Each of these schools had an intensive hiring process, which ensured that new teachers expected to improve in response to feedback. Also, principals should do their best to ensure that teachers are matched with evaluators who know their content area and can model exemplary practices. Some districts successfully do this by assigning peer evaluators who are responsible for both supporting and assessing colleagues (Papay & Johnson, 2012), an option included in the Massachusetts policy, but not part of these schools' approaches. For their part, teachers can step up to new leadership roles as they become available.

**Research.** Recent research has found that combining multiple measures of teachers' performance yields more valid assessments of their effectiveness (Bill and Melinda Gates Foundation, 2013). Similarly, research about policy implementation could benefit from relying on multiple methods and theoretical frameworks to investigate and explain what contributes to good policy and effective policy implementation. If research is to inform policymakers and practitioners, then it must investigate and report much more about how policies are implemented within schools. This can be done by drawing upon teacher surveys, administrative data about employment, and records of day-to-day activities, which together can provide a rich account of policy implementation and effects.

This study yields findings about what works in promoting evaluation for development in a set of successful schools. It would be worthwhile to conduct similar, fine-grained studies in different types of schools, for example, those where principals have little control over hiring and assignment, or where the evaluation policy requires that student achievement constitute a fixed percentage of every teacher's rating. By analyzing particular policies in context, we can come to understand and then explain how implementation happens at the "street level," where students are most directly affected.

Teacher evaluation policies currently leave much unspecified. They set the basics, such as specifying elements of the assessment tool, requiring announced and unannounced observations, and stating whether and how student achievement is incorporated into summative ratings. Within those boundaries, many outcomes are possible. This flexibility may accommodate variation in local needs and priorities, with some schools focusing implementation on accountability and others on development. In deciding how policies should be written, implemented, and monitored, state officials can benefit from knowing much more about how these play out in schools. Meanwhile, district and school administrators need to know much more about how best to achieve desired policy goals in evaluation—to develop the teachers they have, to inform employment decisions, and to skillfully combine both.

# References

Bill and Melinda Gates Foundation (2013). Feedback for better teaching: Nine principles for using measures of effective teaching.

Chester, M.D. (2011, April 16). Memorandum to Members of the Board of Elementary and Secondary Education. Proposed regulations on evaluation of educators.

Coburn, C. E. (2001). Collective sensemaking about reading: How teachers mediate reading policy in their professional communities. *Educational Evaluation and Policy Analysis*, *23*(2), 145-170.

Coburn, C. E. (2005). Shaping teacher sensemaking: School leaders and the enactment of reading policy. *Educational Policy*, *19*(3), 476-509.

Cohen, D. K. (1990). A revolution in one classroom: The case of Mrs. Oublier. *Educational Evaluation and Policy Analysis*, *12*(3), 311-329.

Cohen, D., & Hill, H. (2001). *Learning policy: When state education reform works.* New Haven: Yale University Press.

Cubberley, E.P. (1915). The Portland Survey: A textbook on city school administration. Yonkers-on-Hudson: World Book.

Curtis, R., Weiner, R., & Aspen, I. (2012). Means to an end: A guide to developing teacher evaluation systems that support growth and development. Aspen Institute.

Dobbie, W., & Roland G. Fryer, J. (2011). *Getting beneath the veil of effective schools: Evidence from New York City* (Working Paper No. 17632). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w17632

Doherty, K., & Jacobs, S. (2013). Connect the Dots: Using evaluations of teacher effectiveness to inform policy and practice. National Council on Teacher Quality.

http://www.nctq.org/dmsView/State_of_the_States_2013_Using_Teacher_Evaluations_NCTQ_Report

Donaldson, M. L. (2009). *So long, Lake Wobegon? Using teacher evaluation to raise teacher quality*. Washington, D.C.: Center for American Progress.

Donaldson, M.L., & Casey (2015). Implementing student learning objectives and classroom observations in Connecticut's teacher evaluation system. In Grissom, J.A. & Youngs, P. Eds. *Improving teacher evaluation systems: Making the most of multiple measures*. New York City: Teachers College Press. 131-142.

Donaldson, M. L., & Papay, J. P. (2015). An idea whose time had come: Negotiating teacher evaluation reform in New Haven, Connecticut. *American Journal of Education*, *122*(1), 39-70.

Donaldson, M. L., & Peske, H. G. (2010). *Supporting effective teaching through teacher evaluation: A study of teacher evaluation in five charter schools*. Center for American Progress.

Drake, T. A., Goldring, E., Grissom, J. A., Cannata, M., Neumerski, C., Rubin, M., & Schuermann, P. (2015). Development or Dismissal? Exploring Principals' Use of Teacher Effectiveness Data. In Grissom, J.A. & Youngs, P. Eds. *Improving teacher evaluation systems: Making the most of multiple measures*. New York City: Teachers College Press. 116-130.

Firestone, W. A., Nordin, T. L., Shcherbakov, A., Kirova, D., & Blitz, C. L. (2014). New Jersey's Pilot Teacher Evaluation Program: Year 2 Final Report. New Brunswick, NJ: Rutgers Graduate School of Education.

Institute for Education Science (2014). *State requirements for teacher evaluation policies* *promoted by Race to the Top*. NCEE Evaluation Brief.

Grissom, J. A., & Youngs, P. Editors. (2015). *Improving Teacher Evaluation Systems: Making the Most of Multiple Measures*. New York: Teachers College Press.

Kimball, S. M., & Milanowski, A. (2009). Examining teacher evaluation validity and leadership decision making within a standards-based evaluation system. *Educational Administration Quarterly*, *45*(1), 34–70.

Kingdon, J. W. (2002). *Agendas, alternatives, and public policies*. Boston: Longman.

Kraft, M. A., & Gilmour, A. (2015). Can Evaluation Promote Teacher Development? Principals' Views and Experiences Implementing Observation and Feedback Cycles.

Lipsky, M. (2010). *Street-Level Bureaucracy, 30th Ann. Ed.: Dilemmas of the Individual in Public Service: Dilemmas of the Individual in Public Service*. Russell Sage Foundation.

Massachusetts Task Force on the Evaluation of Teachers and Administrators (2011). *Building a breakthrough framework for educator evaluation in the commonwealth:* Malden, MA: Massachusetts Department of Elementary and Secondary Education.

Maxwell, J. A. (1996). *Qualitative research design*. Thousand Oaks, CA: Sage Publications.

McDonnell, L.M., & Elmore, R.F. (1987). Getting the job done: Alternative policy instruments. *Educational Evaluation and Policy Analysis, 9*(2): 133-152.

McDonnell, L.M., & Weatherford, M. S. (2013). Organized interests and the Common Core. *Educational Researcher,42.* 476-487.

McLaughlin, M.W. (1987). Learning from experience: Lessons from policy implementation. *Educational Evaluation and Policy Analysis, 9*(2). 171-178.

McGuinn, P. (2012*). The state of teacher evaluation reform: State education agency capacity and the implementation of new teacher-evaluation systems.* Center for American

Progress*.*

McGuinn, P. (2015). *Evaluating Progress: State education agencies and the implementation of new teacher evaluation systems.* Consortium for Policy Research in Education.

McLaughlin, M. W., & Pfeifer, R. S. (1988). *Teacher evaluation: improvement, accountability, and effective learning*. Teachers College Press.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Thousand Oaks: Sage Publications.

National Council on Teacher Quality. (2013). *2013 State Teacher Policy Yearbook*.

O'Pry, S. C., & Schumacher, G. (2012). New teachers' perceptions of a standards-based performance appraisal system. *Educational Assessment, Evaluation and Accountability*, *24*(4), 325–350.

Papay, J.P., & Johnson, S.M. (2012). Is PAR a good investment? Understanding the costs and benefits of teacher peer assistance and review programs. *Educational Policy, 26*(5), 696- 729.

Pressman, J.L. & Wildavsky, A. (1973). *Implementation: How great expectations in Washington are dashed in Oakland.* Berkeley, CA: University of CA Press.

Reinhorn, S. K., & Johnson, S. M. (2014). Can evaluation provide both accountability and development for teachers? Evidence from six high-poverty schools. Project on the Next Generation of Teachers, working paper.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417–458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *The American Economic Review*, *94*(2), 247–252.

Sartain, L., Stoelinga, S. R., & Brown, E. R. (2011). *Rethinking teacher evaluation in Chicago: Lessons learned from classroom observations, principal-teacher conferences, and district implementation*. Consortium on Chicago School Research.

Spillane, J. P., Reiser, B. J., & Reimer, T. (2002). Policy implementation and cognition: Reframing and refocusing implementation research. *Review of Educational Research*, *72*(3), 387-431.

Spillane, J. P., Reiser, B. J., & Gomez, L. M. (2006). Policy implementation and cognition: The role of human, social, and distributed cognition in framing policy implementation. *New directions in education policy implementation: Confronting complexity*, 47-64.

Steinberg, M.P., & Donaldson, M.L. (in press). The new educational accountability: Understanding the landscape of teacher evaluation in the post-NCLB era. *Education Finance and Policy.*

Taylor, E. S., & Tyler, J. H. (2012). The effect of evaluation on teacher performance. *The American Economic Review*, *102*(7), 3628–3651.

Taylor, K. (2015, March 22). Cuomo fights rating system in which few teachers are bad. *The New York Times*. Retrieved from http://www.nytimes.com/2015/03/23/nyregion/cuomo-fights-rating-system-in-which-few-teachers-are-bad.html

Thomas, E., Wingert, P., Conant, E., & Register, S. (2010). Why we can't get rid of failing teachers. *Newsweek*, *155*(11), 24–27.

Toch, T., & Rothman, R. (2008). *Rush to judgment: Teacher evaluation in public education*. Education Sector.

von Frank V. (2013) *Evaluations serve as pathways for professional growth.* Learning

Forward.

Weingarten, R. (---). *Teacher development and evaluation: Washington, D.C.* American Federation of Teachers. http://www.aft.org/position/teacher-development-and-evaluation

Weisberg, D., Sexton, S., Mulhern, J., Keeling, D., Schunck, J., Palcisco, A., & Morgan, K. (2009). *The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness. Second Edition*. New Teacher Project.