*Technical Report*

# EVALUATING TEACHERS IN THE BIG EASY: HOW ORGANIZATIONAL CONTEXT SHAPES POLICY RESPONSES IN NEW ORLEANS

EDUCATION
RESEARCH ALLIANCE
FOR NEW ORLEANS

Julie A. Marsh, Susan Bush-Mecenas, Katharine O. Strunk, University of Southern California
Jane Arnold Lincove, University of Maryland-Baltimore County
Alice Huguet, Northwestern University

*April 27, 2017*

EducationResearchAllianceNOLA.org

*Educational Evaluation and Policy Analysis*
*Month 201X, Vol. XX, No. X, pp. 1–32*
*DOI: 10.3102/0162373717698221*
*© 2017 AERA. http://eepa.aera.net*

# Evaluating Teachers in the Big Easy: How Organizational Context Shapes Policy Responses in New Orleans

**Julie A. Marsh**
**Susan Bush-Mecenas**
**Katharine O. Strunk**
*University of Southern California*
**Jane Arnold Lincove**
*University of Maryland, Baltimore County*
**Alice Huguet**
*Northwestern University*

*Although multiple-measure teacher evaluation systems have gained popularity in the United States, few studies have examined their implementation or how they are shaped by organizational context. New Orleans provides a strategic case to examine the enactment of a state teacher evaluation policy in a highly decentralized setting with variation in organizational context. Utilizing a multiple case study approach, we analyzed documents and interviews in eight case study schools. We found that schools varied in their responses to teacher evaluation—in ways that were reflective, compliant, and/or distortive—and that the type of response was not associated with governance model, school authorizer, or level of autonomy. Instead, shared instructional leadership and structures for frequent collaboration appeared to facilitate more reflective responses.*

Keywords: *teacher evaluation, policy implementation, administrators, organizational learning*

In the past decade, multiple-measure teacher evaluation systems (MMTES) have rapidly gained popularity in states and districts nationally. MMTES typically consist of an observation-based measure of teacher effectiveness, a measure of teachers' contributions to student achievement, and often other assessments of teachers' practice (e.g., stakeholder surveys). The objective of MMTES is to provide rigorous and targeted information about teacher performance to help teachers improve their practice and administrators manage the teaching workforce. Spurred in large part by federal programs such as Race to the Top (RTTT) and No Child Left Behind (NCLB) waivers—which incentivized performance-based educator evaluation systems in the criteria for receiving grants or waivers—by 2015, all 50 states and the District of Columbia had policies requiring teacher evaluation and 43 of them mandated the consideration of student achievement data in these evaluations (Doherty & Jacobs, 2015).

As the federal Every Student Succeeds Act (ESSA) ushers in a new accountability era, allowing states greater flexibility around teacher evaluation, it is important not only to take stock of the efficacy of MMTES, but also to understand how MMTES have been implemented on the ground. Extant literature has primarily focused on either the reliability and validity of evaluation measures or the efficacy of evaluation systems at improving teacher effectiveness. Less attention has been paid to school-level implementation and the ways in which organizational context shapes such reforms in practice. As the next generation of evaluation policy takes shape with potentially expanded options for state and district variation, it is helpful to understand how characteristics of educational organizations may affect the implementation and outcomes of policy choices.

Interestingly, whereas federal and state policy have moved toward centralized systems of teacher evaluation, urban school districts have moved toward decentralized control of many components of school operation. This is evidenced by the increasing presence of charter schools and autonomous public school models in cities such as Los Angeles, New York, Chicago, Denver, Detroit, and New Orleans. It is unclear how decentralization of school control facilitates, subverts, or otherwise mediates the implementation of state policy—an issue that will be increasingly important as school autonomy grows and spreads.[1]

New Orleans provides a strategic case to examine the implementation of MMTES in a highly decentralized local setting with wide variation in organizational context. In New Orleans today, a small number of schools are district-run whereas the majority are either single-site charters or operated by charter management organizations (CMOs). Despite this state-driven focus on decentralization of school control, in 2010, the legislature also mandated the use of a new state MMTES, which they named Compass, in almost all schools in the state—including charters.

This standardized system, superimposed upon New Orleans's decentralized setting, allowed us to examine the following questions:

**Research Question 1:** How and to what extent does the design and implementation of state-driven evaluation policy vary across school settings?

**Research Question 2:** What organizational factors are associated with variation in school implementation?

Drawing on qualitative data from eight case study schools, including both traditional and charter schools, we find substantial variation in the implementation of Compass at the school level. To frame our analysis, we draw on concepts from organizational theory and a typology of policy responses adapted from school accountability literature—namely responses that are reflective, compliant, or distortive. After situating our eight case schools within this typology, we examine how these classifications relate to school organizational characteristics.

In what follows, we first describe extant research on teacher evaluations, then the context of New Orleans and the Compass system, followed by a description of our conceptual framework and methods. Next, we discuss our findings regarding the varied design and implementation of Compass across case schools, including how cases demonstrated reflective, compliant, and distortive responses to evaluation, and the organizational factors that seemed to influence responses. We conclude with implications for future policy and research on MMTES across varied contexts.

## Putting the New Orleans Case in Context: What We Know About Teacher Evaluation

In recent years, MMTES has received a great deal of attention from policymakers (Lee, 2010; U.S. Department of Education Office of Inspector General, 2011), researchers (e.g., Ellett & Teddlie, 2003; Weisberg, Sexton, Mulhern, & Keeling, 2009), and the popular press (e.g., Baker, 2013; Kenny, 2012; Leonhardt, 2013). Existing research has examined the reliability and validity of evaluation measures (e.g., Chetty, Friedman, & Rockoff, 2014a, 2014b; Hill, Charalambous, & Kraft, 2012; Hill, Kapitula, & Umland, 2011; Ho & Kane, 2013; Kane & Staiger, 2012; Kane, Taylor, Tyler, & Wooten, 2011; Kimball, 2002; Sartain, Stoelinga, & Brown, 2009; Steinberg & Garrett, 2016; Strunk, Weinstein, & Makkonnen, 2014; Taylor & Tyler, 2012) and the efficacy of evaluation programs (e.g., Dee & Wyckoff, 2013; Taylor & Tyler, 2012).

A limited body of research has investigated educators' experiences implementing MMTES. These studies often indicate that, although teachers and administrators hold generally positive views of new evaluation systems (Donaldson et al., 2014; Hamilton et al., 2014; Jiang, Sporte, & Luppescu, 2015), they have also expressed concerns. Most implementation studies demonstrate that new evaluation systems require a significant time commitment from principals, which is often perceived as a burden (Donaldson &

Cobb, 2015; Heneman & Milanowski, 2003; Kimball, 2002; Milanowski & Heneman, 2001; Sartain et al., 2009) and an obstacle to conducting multiple observations and conferences, particularly in the early years of implementation (Derrington & Campbell, 2015; Donaldson et al., 2014; Murphy, Hallinger, & Heck, 2013; Sartain, Stoelinga, & Brown, 2011). Such challenges are likely to influence how teachers and administrators respond to policy.

Studies also have uncovered challenges implementing the observation and feedback processes expected in MMTES, and in particular those related to individual and organizational capacity. Research indicates that evaluators are often unwilling or unable to identify low-performing teachers (Strunk et al., 2014; Tennessee Department of Education [TDOE], 2012) and that they struggle to differentiate among teachers producing moderate student achievement results (Jacob & Lefgren, 2008). Other research identifies weaknesses in the quality of feedback. Conversations between principals and teachers regarding observation- and student performance–based evaluations are often focused on classroom management rather than instructional methods, not sufficiently tailored to the teacher's specific subject-matter, and dominated by basic questions and evaluator talk rather than meaningful guided reflection by the teacher (Heneman & Milanowski, 2003; Kimball, 2002; Milanowski & Heneman, 2001; Sartain et al., 2011). In addition, administrators lack the needed preparation to implement evaluation systems and would benefit from additional training on coaching skills (Bell et al., 2012; Derrington, 2014; Donaldson et al., 2014; Sartain et al., 2011).

Despite these challenges, there is some promising evidence that MMTES have in particular cases encouraged teachers and principals to engage in more reflective conversations regarding their practice (Heneman & Milanowski, 2003; Kimball, 2002; Sartain et al., 2009; Strunk et al., 2014). In their study of Connecticut's new evaluation policy, Donaldson and Cobb (2015) found that teachers benefited from receiving more feedback than under previous evaluation systems. Similarly, in a study of superintendents and principals in four districts, Derrington (2014) found consensus that an MMTES led principals to improve their abilities to support teachers' instructional improvement. A study of an MMTES in five charter schools also found that the new system promoted professional growth by focusing on a goal of continuous improvement, developing a reflective culture, and regularly involving teachers in discussions about their performance (Donaldson & Peske, 2010). While indicating the potential for MMTES to spur improvements in practice, these studies do not explicitly identify the conditions necessary to do so.

Although extant research highlights the promise and challenges of implementing MMTES, there is a dearth of research on how local evaluation design, school governance, and organizational context shape implementation. Our study seeks to build a deeper understanding of the quality of implementation and the factors shaping local responses to state evaluation policy in New Orleans.

### *The Case of New Orleans and Its Differentiated School Organizations*

New Orleans is an ideal case to examine how a standardized state evaluation policy plays out across various school contexts. Following the citywide evacuation and destruction of many school buildings during Hurricane Katrina, the New Orleans public school system was radically reformed as a decentralized system composed mostly of independent charter schools. This process began in 2005–2006 when, following the temporary shutdown of all schools due to the evacuation, the state's Recovery School District (RSD)[2] took over all underperforming public schools in the city. As a severely underperforming school district, this left only a small number of historically high-performing schools in the hands of the locally elected Orleans Parish School Board (OPSB). Between 2006 and 2014, RSD either permanently closed or contracted all of the schools under its control to nonprofit charter operators. By 2013–2014, the year of this study, more than 90% of New Orleans public school students attended charter schools (Louisiana Department of Education [LDOE], 2015).

New Orleans charter schools enjoy a high level of autonomy over many school features including grades offered, school days and hours, instructional strategies, and teacher hiring, compensation, and professional development (PD). This is particularly important given the high-risk, historically underserved population in most public schools—New Orleans's public school students are 95% Black and 85% are eligible for free and reduced-price lunch. The combined effects of Hurricane Katrina and school reform have significantly reduced teacher average experience from 15.2 years (2005) to 9.1 years (2014) and increased teacher turnover rates from 9.9% (2004) to 17.9% (2013).[3] Thus, teacher quality and improvement are a central concern in the system overall. However, most state laws relating to teachers (e.g., teacher certification requirements) do not apply to charter schools. In fact, Compass, described below, is an unusual case of a Louisiana state education policy that was adopted for implementation in both charter and traditional public schools.

Within this system, New Orleans has a diverse group of schools. Three public entities are responsible for charter contracting—the RSD, OPSB, and the Louisiana Board of Elementary and Secondary Education (BESE).[4] Several different types of schools operate in New Orleans, including traditional schools overseen by OPSB, charters in a CMO network, and single-site charters. The different types and layers of management and governance (or lack thereof) may affect the level of discretion experienced by and resources available to school-level educators, particularly in relation to the implementation of policy mandates. Compared with districts in which teacher evaluation is subject to district-union negotiations, the absence of a teacher collective bargaining agreement in New Orleans (even for teachers in traditional public schools) may also facilitate variation in personnel policies and school operations. In addition, almost all New Orleans charter and district-run schools offer open enrollment and compete for students in a citywide enrollment system. Thus, a drive for innovation and differentiation might facilitate high levels of variation (Arce-Trigatti, Harris, Jabbar, & Lincove, 2015). This atypical variation in school governance, management, and programs provides an important opportunity to explore how teacher evaluation policy is implemented in differentiated contexts.

### *The Compass System*

To better understand the design and objectives of Compass, we interviewed state officials and reviewed state laws and policy documents (for more detail, see our discussion of methods). We traced the origins of Compass to 2010 when, according to a former staff member for then-governor Bobby Jindal, statewide concerns about the quality of teaching were percolating in the governor's office. As a response to these calls for accountability and in pursuit of an RTTT grant, Louisiana's legislature passed Act 54, which mandated an expanded teacher evaluation system. Although they were not awarded an RTTT grant that year—they did receive one in 2011—Louisiana moved forward in developing the proposed evaluation system.

Act 54 of 2010 mandated that all Louisiana public school teachers receive an annual evaluation consisting of two equally weighted components: measures of student performance growth and observations of teaching. Rather than prescribe particular methods or tools for determining these measures, Act 54 called for an appointed Advisory Committee on Educator Evaluation (ACEE) to recommend strategies for the development of value-added measures and standards of teacher effectiveness. In 2011, ACEE presented their recommendations to BESE in the form of revisions to state regulations (Bulletin 130), which were the foundation for the evaluation system now known as Compass (ACEE, 2011).

According to an LDOE official interviewed for this study, the goal of Compass was "to make sure that we elevate the quality of teaching" and "increase student achievement as a result of this process." The mechanisms for improvement, as conveyed in policy documents, were twofold. Enhanced feedback on teacher practice and student performance was intended to help teachers reflect on and improve instruction, and information gathered in the process was intended to inform staffing decisions, ranging from assignments to termination.

Aside from a few requirements, Compass placed control in the hands of local administrators, leaving LDOE to monitor compliance and ensure that Local Education Agencies (LEAs) and CMOs submit evaluation ratings. As one LDOE official said,

> I think what's unique about our state is that we really, truly believe that the school leader is the best person to use this tool to achieve results, and that the state's role is not to limit the principal's authority to work with their teachers to improve.

Bulletin 130 specified that teachers' Compass ratings must include measures of student growth and classroom observations. When available, the student-growth factor in a teacher's evaluation was the state-calculated valued-added measure (VAM) based on state standardized tests for teachers of tested grades and subjects. For other teachers, the growth measure used Student Learning Targets (SLT) selected and measured at the school level. Although SLTs were to be based on state-approved common assessments where available, other measures such as student portfolios could be used in subjects without standardized assessments. According to one LDOE official, principals were "charged with evaluating the quality of an SLT, gauging end-of-year attainment of the target, and submitting components of the evaluation to the state," which allowed for flexibility across schools. Bulletin 130 also gave LEAs the ability to define learning targets across similar classrooms. According to an LDOE official, Compass' SLT process was intended "to be an authentic exercise that a teacher and a principal go through to understand where students are performing, what a reasonable but ambitious goal is, and to set a goal based on that." Important to our study, in 2014, Louisiana transitioned from state-developed standardized tests to the new Common Core–aligned test. In the initial year of new state testing—also the year of our study—teachers did not receive test-based VAMs and all Compass growth measures were based on SLTs.

According to Bulletin 130, the other half of a teacher's Compass rating had to be derived from classroom observations conducted by principals, assistant principals, or other designees who obtain evaluator certification through a 2-day training provided by LDOE. Teachers were to be observed a minimum of two times during the school year, including one "unannounced" visit and one prescheduled "announced" observation, each lasting one full class period. The announced observation was preceded by a preobservation meeting between the teacher and observer, and followed by a postobservation conference. Using an abbreviated rubric adapted from Charlotte Danielson's Framework for Teaching (Danielson, 2013), observations were assessed on a 4-point scale in three domains: planning and preparation, classroom environment, and instruction (LDOE, 2013). In response to feedback that "many administrators struggled to offer teachers frequent, authentic feedback on a tool that was that large," one state leader explained, policymakers selected just five of Danielson's 76 rubric elements to include in the Compass rubric: (a) setting instructional outcomes, (b) managing classroom procedures, (c) using questioning and discussion, (d) engaging students in learning, and (e) using assessment in instruction. As an LDOE official reported, "We essentially analyzed the rubric to make a decision about the components that were most related to success with students." Notably, the text of the selected elements primarily emphasizes high levels of student engagement.

Bulletin 130 also gave LEAs—including charter school operators—the option to develop or identify their own observation tools in lieu of the state-provided rubric by submitting a waiver and justification to the state. Officials developed this exception, in part, because they recognized many charter schools already had similar evaluation systems in place. LDOE officials estimated that about one third of LEAs and CMOs used an alternative observation tool at the time of our study.[5] Furthermore, Bulletin 130 stipulated that, to preserve due process, charter schools were not required to follow certain provisions, such as who completes evaluation, how support is provided to teachers, and the grievance process. These modifications, however, did not change the intent of the policy, which was to create a consistent teacher evaluation process across district and charter schools. This intent was confirmed in our case findings, as staff perceived the Compass policy as applying to all school types and fully binding to charter schools.

In the end, LDOE was expected to aggregate the two scores from student growth and observations into a final composite score, weighting each component equally. Final scores were sorted to one of four categories: ineffective, effective emerging, effective proficient, or highly effective. Importantly, a teacher could be rated

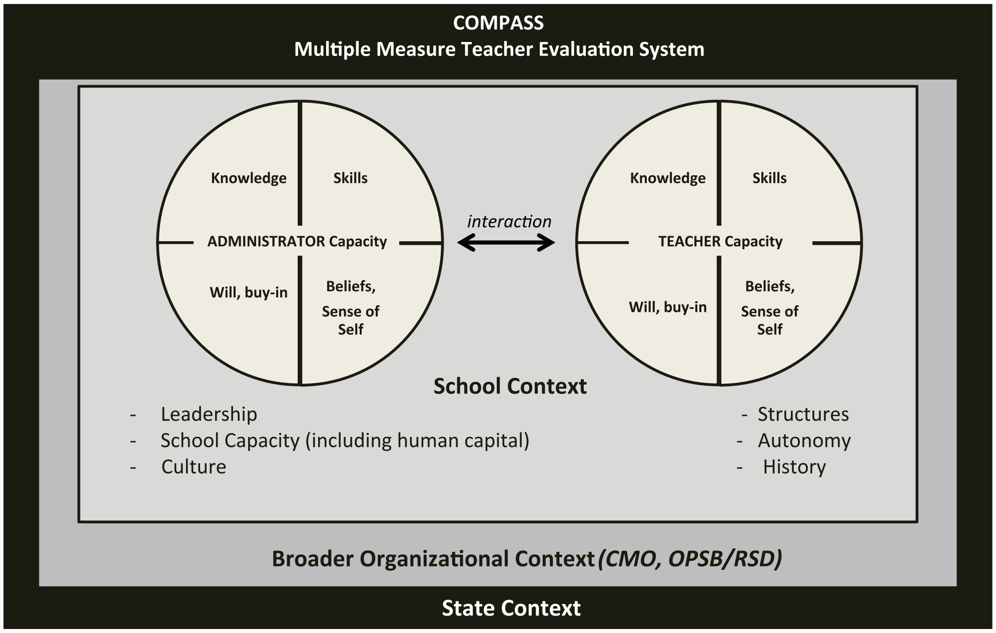

FIGURE 1. *Conceptual framework.*
*Note.* CMO = charter management organization; OPSB = Orleans Parish School Board; RSD = Recovery School District.

ineffective due to an ineffective rating on either the growth or observation measures, as well as if the aggregate score was ineffective. If teachers were rated ineffective, schools were required to work with them to develop an intensive assistance plan. If a teacher was rated ineffective more than once, the LEA had to initiate termination proceedings within 6 months (for charter schools, teachers receiving ineffective ratings for 3 consecutive years had to be terminated). Yet the asserted purpose of Compass was not punitive. Rather, Bulletin 130 required LEAs to provide PD aligned with teachers' individual Compass-identified areas of growth to generate measurable improvements in teacher practice.

## Conceptual Framework

We draw on two bodies of literature to guide our analysis of Compass implementation: studies of educational accountability-policy implementation and organizational theory. As depicted in our integrated conceptual framework (Figure 1), we begin at the core of learning in teacher evaluation—the capacity of teachers and administrators and their social relations. Through interaction between principals and teachers (the center reciprocal arrow in Figure 1), administrators observe instruction and formulate scores on the observation rubric, and teachers and administrators negotiate SLTs and discuss teacher performance and improvement opportunities, all of which may spark learning. The organization as a whole may also learn through this process, as administrators and teachers garner greater information about schoolwide practice, enabling adjustments to facilitate improvement. Individual learning is situated within the school, broader organizational (such as the CMO or central office), and state context, considering various organizational factors derived from the literature (discussed below). Finally, we overlay the concept of individual and organizational responses to teacher evaluation, which may manifest in reflective, distortive, or compliant ways.

### *Responses*

To categorize the types of school responses to Compass, we draw on organizational theory and empirical literature. Theories of organizational learning suggest that organizations engage in a range of practices when called upon to use information, and that practices are likely modified over time in light of feedback from the environment (Scott, 1998). Scholars note that learning can at times be "productive" and at other times "limited, distorted, and misdirected" (Smylie, 2009, p. 33; see also, Levitt & March, 1988; March, 1994).

Consistent with other studies examining education policy implementation through the lens of organizational learning (e.g., Coburn, 2001, 2005; Honig, 2003, 2012; Spillane & Miele, 2007; Spillane, Reiser, & Reimer, 2002), we draw on these basic organizational learning precepts to understand how New Orleans schools respond to new state demands for teacher evaluation and utilize the new information about teacher practice that Compass provides. In particular, we build on Jennings's (2012) application of such a lens to teachers' use of student achievement data, in which she defines "productive data use" as "practices that improve student learning and do not invalidate the inferences about student- and school-level performance that policy makers, educators, and parents hope to make" (p. 4). Conversely, she characterizes "distortive data use" as

> use of test score data to make instructional and organizational decisions produc[ing] score gains that do not generalize to other measures of learning . . . and thus lead[ing] [teachers] to make invalid inferences about which schools, teachers, and programs are effective. (p. 4)

In fact, many studies find that schools and educators respond to performance measures and accountability policy in distortive ways, adopting practices that boost test scores and a school's chances of reaching proficiency targets (e.g., focusing on "bubble kids") rather than genuine improvement and learning (Booher-Jennings, 2005; Hamilton et al., 2007; Mintrop, 2012; O'Day, 2002).

Building on this body of literature and emergent patterns of implementation observed in our early data collection and analysis, we defined three types of school-level responses to teacher evaluation policy: reflective, distortive, and compliant. Schools may respond in *reflective* ways, engaging in reportedly meaningful reflection (producing and using evaluation data to think about instruction and ways to improve it) and, in some cases, taking actions to bring about improvement. Schools with reflective responses generally perceive evaluation data as valid measures of teaching to inform improvement efforts.[6] Other schools may respond in *distortive* ways, which preclude reflection. This could mean taking actions that lead to actual or perceived invalid measures of teaching practice, such as engaging in strategic behaviors to appear effective according to evaluation criteria. Finally, we expand Jennings's (2012) typology by acknowledging that schools may exhibit a *compliant* response, following technical requirements but not embracing "the spirit" of the policy (McLaughlin, 1987). Such schools may "go through the motions," but not reflect or act to change practice. We return to this typology and how we operationalized it in the "Data and Method" section.

### *Nested Organizational Context*

Decades of policy implementation research point to a pattern of variation in the implementation of school reforms (e.g., Honig, 2012; McLaughlin, 1987) and the ways in which context shapes policy implementation (e.g., Louis, 2007; McLaughlin & Talbert, 1993; Spillane & Louis, 2002). Broader organizational theory and empirical literature on schools identifies a host of organizational factors likely to shape school-level responses, as illustrated in the inner gray boxes in Figure 1 (individual capacity is captured in the teacher and administrator circles).[7]

*School Context.* Theory and research suggest a host of interrelated, school-level contextual factors that may affect local responses to MMTES. First, *leadership*—both leadership style and the distribution of leadership within the organization—may play an important role in how schools learn and act on new policy demands. Leaders are thought to play an important role in shaping organizational learning by providing (a) continuous challenges to members, (b) freedom to innovate, (c) resources to innovate, (d) diverse perspectives within teams, (e) encouragement, and (f) support (Amabile, 1997). Leaders are crucial in encouraging learning and innovation among members (Shallcross, 1975; Suh, 2002), and shared leadership holds great potential for promoting empowerment, a growth mind-set, and learning within organizations (Goldsmith, Morgan, & Ogg, 2004). Of course, shared leadership depends upon strong human capital management, to ensure that members possess the requisite skills and dispositions to engage in continuous learning (Shipton, Dawson, West, & Patterson, 2002). Human capital reforms, particularly MMTES, place particular demands on leaders, as they rely upon complex interactions among organizational members. Studies of leadership in schools indicate that administrators are integral to reform and may influence student learning indirectly by structuring the school organization, culture (Waters, Marzano, & McNulty, 2003), and teacher working conditions (Leithwood, 2006; McLaughlin & Talbert, 2001), and by interacting with teachers (Cuban, 1988; Hallinger & Leithwood, 1994; Heck, 1993) and facilitating reflection (Blase & Blase, 1999). Furthermore, studies find that *instructional leadership*—supporting and holding teachers accountable for high-quality instruction for all students—positively affects teaching and learning (e.g., Hitt & Tucker, 2016; Leithwood, Louis, Anderson, & Wahlstrom, 2004; Purkey & Smith, 1983; Robinson, Lloyd, & Rowe, 2008).

*Organizational capacity*—the ability of an organization to fulfill its mission and goals—is another important factor to consider. Organizational capacity is informed in part by the skills, knowledge, and experience of individuals (human capital) as well as social capital (networked relationships among staff), program coherence (integration of instruction, resources, and staff), and resources (Beaver & Weinbaum, 2012; King & Bouchard, 2011; Newmann, King, & Youngs, 2000). Furthermore, years of empirical research indicate that education policy implementation depends upon the capacity and will of educators (McLaughlin, 1987). In the context of MMTES, organizational capacity, manifest in peer and leader coaching and access to PD, plays an important role in facilitating human capital improvements.

*School culture*—norms, routines, values, rituals, and expectations, particularly around staff-to-staff, staff-to-student, and student-to-student interactions—also likely affects responses to MMTES. According to some theoretical literature, organizational learning is more likely to occur within organizational cultures that promote innovation and value creativity (Amabile, 1997; Woodman, Sawyer, & Griffin, 1993). One recent study found that aspects of school culture, including respect among teachers and high academic expectations for students, promoted teachers' capacity for organizational learning (Louis & Lee, 2016). School culture may influence expectations around individual improvement efforts, such as motivating teachers to meaningfully implement MMTES.

Leadership and culture closely relate to another important dimension of school context, that of *structures*. School structures include the defined roles and responsibilities within an organization, how tasks are allocated, and how information flows. Organizational structure is inextricably linked to organizational learning (Fiol & Lyles, 1985), as structure defines how processes and people interact (Chen & Huang, 2007; Dodgson, 1993), information is shared (Lloria, 2007), and learning activities are coordinated (Dodgson, 1993). In the educational context, structures that promote organizational learning might include planned teacher collaboration time, interdisciplinary committees, team teaching, and regularly scheduled PD opportunities (Leithwood, Leonard, & Sharratt, 1998). These structures likely influence the communication and understanding of MMTES policy, the subsequent enactment of observations, and the provision of support.

Also, *autonomy*—the extent to which the school organization can make decisions regarding instruction and operations—may drive variation in implementation of MMTES. Extant empirical research indicates that autonomy allows schools flexible hiring processes to recruit high-quality staff committed to their mission (Burian-Fitzgerald, Luekens, & Strizek, 2004; Gross, 2011) and modify school structures that facilitate teacher collaboration and student support (Doyle & Feldman, 2006). Thus, autonomy around staffing and structures may enable schools to engage in purposeful learning, informed by their particular context, history, and survival pressures. This flexibility, in theory, may facilitate more reflective responses to evaluation policy, providing administrators leeway to organize time for sharing feedback and PD to respond to identified areas of need.

Finally, *history*, described as the ability to retain useful practice, knowledge, and learning abilities, and to unlearn ineffective practices, also plays an important role in guiding organizational learning (El Sawy, Gomes, & Gonzalez, 1986; Lane & Lubatkin, 1998). In the educational setting, history can be a function of school structures, norms, programs, and organizational capacity (in this case, including the LEA or school's past experience with teacher evaluation).

*Broader Organizational Context.* Next, we examine the influence of other organizational factors—related to school governance, authorizer, and management—on organizational learning through MMTES across diverse school settings, as illustrated in the darker gray box in Figure 1. In New Orleans, some RSD charters are managed by large CMOs that operate up to six schools under a single governing board, whereas many are single-site charters. Furthermore, all RSD charter schools are former underperforming district schools that were subject to state takeover, followed by contracting to a charter operator. OPSB charters include historically high-performing schools that voluntarily transitioned to charter status and other new schools that have opened since 2006.

This variation allows us to examine how organizational factors influence responses to state evaluation policy. Farrell, Wohlstetter, and Smith (2012) posit that CMO networked schools may be well positioned to implement educational reforms as they are more nimble than school districts (allowing them to innovate and disseminate "best practices") while having greater leverage and capacity to support these activities than single-site charters—although the empirical findings on this topic are mixed (e.g., Lubienski, 2003; Preston, Goldring, Berends, & Cannata, 2012). For example, the less complex organizational structure of CMOs (compared with traditional districts) may allow principals to focus on being instructional leaders, whereas leaders in single-site charter schools may experience demands on their time to address facilities, budgets, and so on (Cravens, Goldring, & Penaloza, 2012). However, it is also possible that unlike stand-alone charters, CMOs develop more standardized policies that can limit autonomy of school educators (Lake, Dusseault, Bowen, Demeritt, & Hill, 2010). Thus, the broader organizational context (including aspects such as leadership, capacity, culture, structures, autonomy, and history) may influence the school's local organizational context in ways that influence implementation of and responses to teacher evaluation. In what follows, we detail our data collection and analytic methods, with special attention to how we operationalized the response type and organizational factor constructs.

## Data and Method

We utilized a multiple, exploratory, embedded case study approach (Yin, 2013), including a purposeful sample of eight case study schools to represent variation in governance (traditional vs. charter)—and with charters, further variation by type (single-site vs. networked) and authorizer (RSD vs. OPSB)—and grade levels served (see Table 1). Our sampling logic was theoretical in nature, assuming that school responses to teacher evaluation policy would vary based on differences in organization type. We intended to visit two schools in each of four categories: OPSB direct-run, OPSB CMO charter, RSD CMO charter, and RSD single-site charter. Within each category, we randomly selected two schools to contact, as well as a set of back-ups. In the end, 8 of 15 schools contacted were able to accommodate our visit (2/3 OPSB direct-run, 2/3 OPSB CMO charters, 2/5 RSD CMO charters, and 2/4 RSD single-site charters). Although our sampling and case study design relied on the assumption that the implementation of evaluation programs would vary across schools with different organizational structures, our inductive analyses unearthed more nuanced patterns between organizational factors, rather than school type, and the kind of response to evaluation.

In 2015, we conducted semistructured interviews with LDOE and CMO administrators ($n$ = 3), school administrators ($n$ = 17), and teachers ($n$ = 36). At each school, we requested to speak with one principal, another school leader, four core-subject teachers (four teaching Grades 3–5 for elementary, and two math and two English for secondary), and one teacher in another grade level/subject (see Table 1 for detail).[8] Our interviews covered topics such as the school context (e.g., mission, culture, autonomy, hiring practices), the school's teacher evaluation process (e.g., history, purpose, definition of quality teaching, goal-setting process, use of rubrics, observation process, consequences/incentives),

TABLE 1

*Range of Case Authorizer, Governance Model, Size, Demographics, and Achievement*

| | Range of cases (8 total) |
|---|---|
| Authorizer | RSD (4), OPSB (4) |
| Governance model | Direct-run (2), independent charter (2), CMO charter (4) |
| Grade level | K–8 (5), 9–12 (3) |
| Size | 160–980 students (average = 600) |
| Achievement[a] | A (2), B (3), C (1), F (2) |
| Student demographics | <5%–18% Special Education (average = 11%)<br>65%–95% Economically Disadvantaged (average = 83%)<br>52%–100% African American (average = 88%) |
| Case respondents | Number of interviewees per case<br>3–6 teachers (average = 4.25)<br>0–2 principals (average = 0.875)<br>0–2 other administrators (average = 1.375)<br>Teacher years of experience<br>13%–33% 0–2 years (average = 22%)<br>33%–80% 3–5 years (average = 50%)<br>0%–33% 6–9 years (average = 11%)<br>0%–33% 10+ years (average = 17%)<br>Teacher credential type<br>20%–100% traditional credential (average = 66%)<br>0%–80% alternative credential (average = 27%)<br>0%–33% both credential types (average = 3%)<br>Subject taught<br>50%–86% tested grade/math or English (average = 72%)<br>0%–40% untested/elective/other subject (average = 20%)<br>0%–33% special education (average = 8%) |

*Note.* RSD = Recovery School District; OPSB = Orleans Parish School Board; CMO = charter management organization.

[a] School achievement is presented using Louisiana Department of Education School Report Card Grades (measured from A-F). For more information, see http://www.louisianabelieves.com/assessment/school-letter-grades

and the respondent's last evaluation. All interviews were audio recorded, transcribed, coded, and analyzed (using NVivo qualitative research software). We also reviewed documents collected from case schools (e.g., evaluation rubrics and forms) as well as state documentation regarding Compass (e.g., Bulletin 130, training PowerPoint).

### *Case Analysis*

Through case analysis, we sought to understand (a) how schools implemented Compass, including school-level variation in the design of evaluation systems and how educators responded to them; (b) the organizational characteristics of each case school; and (c) the relationship between implementation and organizational characteristics. First, we analyzed each case individually by writing detailed memos using a standardized template. These initial memos helped to specify the design and implementation of teacher evaluation locally and key contextual elements at each school. We then coded all interview transcripts and relevant documents along the dimensions of Figures 1 and 2. Initially, we coded all interviews according to two sets of codes, determined prior to data collection. One set of descriptive codes, concerning the elements of teacher evaluation (e.g., observations, conferencing, and feedback), was applied to all interviews to help organize data. Another set of thematic codes (drawn from the literature) concerned the organizational characteristics of schools, including history, autonomy, leadership, collaborative structures, culture, and capacity. After data collection, we created a third set of codes to capture the range of responses to teacher evaluation. We used this final set of analytic codes to classify school responses as reflective, compliant, and/or distortive. We compared coding from multiple coders across 15% of our interviews, and interrater reliability fell above a threshold of 70% agreement, considered acceptable for exploratory studies (Campbell, Quincy, Osserman, & Pedersen, 2013; Fahy, 2006). Following coding, we utilized displays to analyze our qualitative data. (See Appendix A for more details, available in the online version of the journal.)

To begin our analyses, we utilized matrix coding functionalities in NVivo software to examine patterns in the frequency of words coded for each response type in each case study. Next, we examined the qualitative data coded under each response type, broken out according to evaluation activities, within each case (for example, we compared compliant responses regarding setting SLTs across cases), with special attention to triangulating data across respondents. Drawing on this detailed analysis, we situated each case according to the overall character of responses to evaluation. Separately, we created a case-ordered descriptive metamatrix (i.e., a table with cases as rows and school characteristics and organizational factors as columns), and used color coding to illustrate different qualities of each case for a construct (e.g., shading cases green, yellow, or red according to the extent to which they reported using shared leadership structures) (Miles, Huberman, & Saldaña, 2013). Finally, we added the overall case response type to our metamatrix to illuminate any patterns among response type and organizational factors (see Appendix B for greater detail, available in the online version of the journal).

We drew upon the quality and quantity (see Appendices B and C, available in the online version of the journal) of coded interview data for each response type to characterize the overall school-level response to teacher evaluation. Concurrently, we populated two sets of matrices: (a) a frequency matrix with the percentage of words coded for each case across each response type (see Appendix C, available at in the online version of the journal) and (b) a content matrix with the qualitative data coded for each case across each response type (see Appendix B, available in the online version of the journal). The frequency matrix enabled us to cleanly assess the quantity of comments tied to different response types, whereas the content matrix shed light on the strength and quality of responses, and consistency across interviewees. Together, these matrices informed how we categorized cases. While we began analysis by conceptualizing responses along a continuum, our analyses demonstrated that schools, individuals, and even statements might demonstrate more than one response type. As opposed to mutually exclusive categories, these responses often overlapped and many schools exhibited more than one type, for example, demonstrating a degree of reflection while utilizing some distortive practices (Figure 2). We utilized this multiple-response categorization when analyzing our data for patterns across response type and organizational factors. (We return to this categorization in the "Findings" section.)

Along the way, researchers wrote memos and met to discuss possible alternative explanations. Together, triangulation and careful coding contributed to construct validity, as we drew on multiple measures of the same phenomenon and provided a clear "chain of evidence" from data to findings (Yin, 2013, p. 186). Furthermore, displays (described above) helped us to see patterns among multiple constructs, and our attention to alternative explanations also helped to ensure the robustness of findings (Yin, 2013).

### *Limitations*

There were several limitations to our data collection and analyses. First, the schools in this study were purposefully sampled and are not meant to be representative of all schools or subgroups of schools in New Orleans. Similarly, we sampled teachers within schools and cannot be sure those interviewed represent typical responses in a given school. Second, our findings are anchored specifically in the unique context of New Orleans, and may not apply to other settings. Third, we did not formally observe school activities, which would grant us additional information regarding the routines, structures, norms, and values that promote organizational learning. We recognize that such measures would provide a deeper understanding of organizational responses to policy, and instead, rely here on interview-based accounts of these factors. Given this shortcoming, we triangulate interview data from multiple sources to ensure the credibility of reports of these organizational factors and practices. Fourth, we rely on one year of data (2015) and thus cannot speak to longer term change or learning over time. Although this study is best understood as exploratory, it is an important first step in examining school responses to state policy mandates in a decentralized context, and provides a framework and initial findings to inform future research.

## Findings

In this section, we provide results from three levels of analysis. First, we describe the teacher evaluation systems implemented in case study schools. As noted, state policy provided districts and charter managers with some flexibility in how they enacted evaluation; thus, understanding the parameters of the evaluation system in each school provides an important context for the second set of analysis. Second, we characterize how teachers and school leaders made use of their local evaluation systems, demonstrating variation in reflective, compliant, and/or distortive responses. Finally, we identify the organizational factors associated with these patterns of response.

### *Variation in School-Level Evaluation System Design*

As expected, case schools varied in how they enacted teacher evaluation systems, with some designing systems that went beyond the state Compass requirements. We observed five important areas of variation: (a) the observation rubric, (b) the number of observations, (c) training to help teacher mentors or coaches assist teachers in developing their practice, (d) guidelines for setting SLTs, and (e) incentives attached to positive ratings.

The basic state model for teacher evaluation under Compass included a rubric made up of five components of the Danielson framework. Two of our eight case schools obtained waivers to use alternative observation rubrics: One used the full Danielson rubric with an altered scale, whereas the other used a school-developed rubric including job-related, instructional, and professional proficiencies. Both of these rubrics were far more detailed and lengthy than the basic state model rubric. Furthermore, under the basic state model, teachers were to be observed twice per year—one announced and one unannounced. In both of the schools with waivers, two additional unannounced observations were added for each teacher. In a third case, teachers receiving an effective rating in their first, announced observation were exempt from the second, unannounced observation. In all cases, administrators reported also conducting shorter, informal observations of instruction, sometimes referred to as "walk-throughs," as a part of their general management practice.

Schools also varied in the ways they prepared observers to engage in evaluation. LDOE required a 2-day observer certification, but did not test rater proficiency or reliability in rating practice, which several school administrators identified as a weakness. In response, several case schools developed additional procedures to enhance the reliability of ratings across observers. At one school, administrators created their own methods of norming, by comparing scores of the same observation between two administrators and developing common understandings of rubric categories and ratings. At another school, teacher content-specialists often accompanied the administrator on observations to offer their expertise (a response to difficulties related to observing content-heavy instruction at the high school level).

All schools reported using standardized (vs. teacher-developed) assessments to set SLTs, but the specific assessments varied widely. Schools reported using a variety of benchmark (DIBELS, Brigance testing, STAR math and STEP literacy assessments) and summative (state End of Course examinations, LEAP, PARCC) assessments. At some schools, administrators determined which assessments would be used, whereas at others, teachers selected from any standardized assessment given. Similarly, some schools required that SLTs cover all students, whereas the majority allowed teachers to set goals and monitor progress for a subset of students. Schools also varied in requirements of who would set performance targets. Although most of our cases used a process of negotiation between the teacher and principal to determine targets (with varying authority afforded to teachers and administrators), in one case, administrators set a specific proficiency target on a standardized test as the SLT for all teachers. As students outperformed the goal, school leaders planned to slightly increase the proficiency target for the subsequent year.

At all sites, educators reported developing personalized improvement plans for teachers rated as ineffective, but we observed substantial variation in performance incentives across sites. In some cases, local evaluation systems included a set of potential consequences, both positive and negative. In five schools, administrators reported counseling out employees who were not making progress. For example, one administrator reported, "I have a saying, 'let me help you out, or let me help you *out*' . . . If the kids are not growing, this is not the place for you." On the reward side, at five schools, highly effective educators were eligible for merit-based bonus pay.

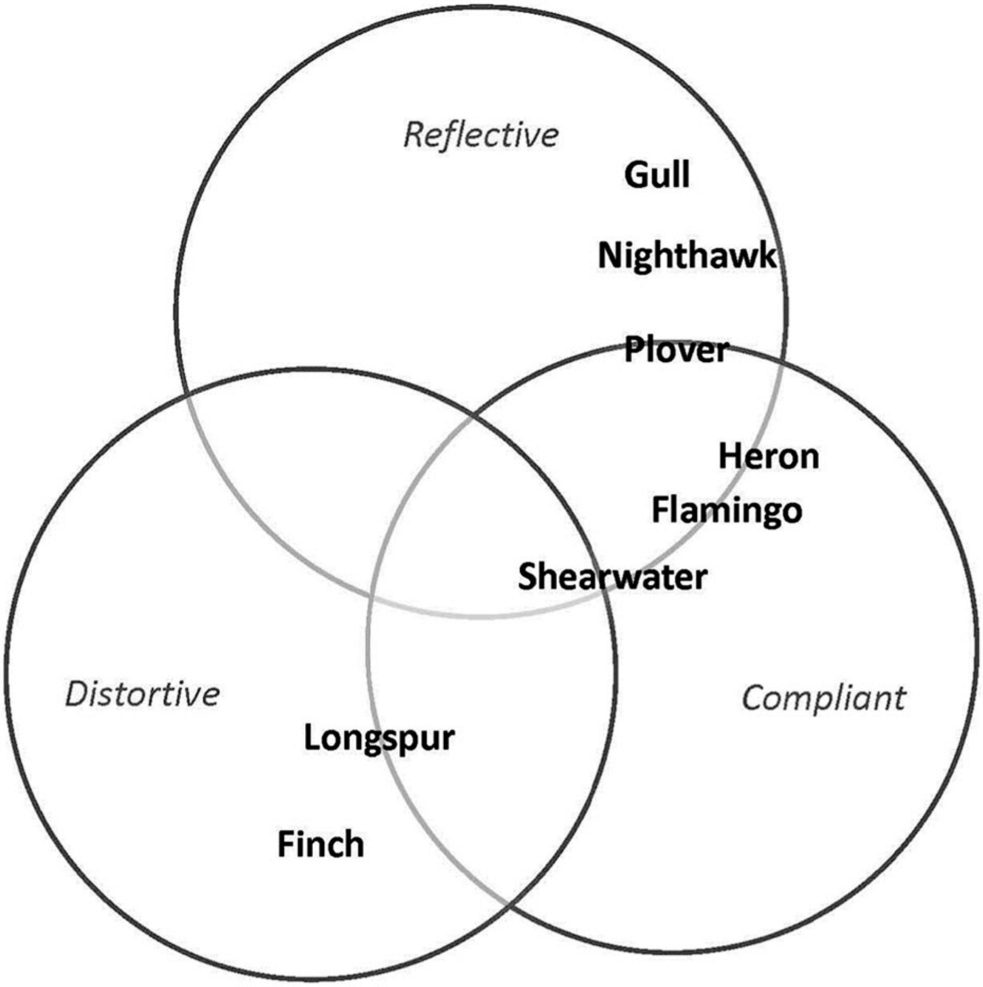

FIGURE 2. *Predominant response types of cases.*

*School-Level Responses to Teacher Evaluation*

In all schools, we found examples of reflective, compliant, and distortive responses to evaluation. Our overall evidence for each case, however, generally indicated either a strong tendency toward one particular response or, in several schools, a split between two categories such that the character of overall case response was classified as reflective-compliant, compliant-reflective, distortive-compliant, compliant-distortive, and even compliant-reflective-distortive (reflecting in order, primary, secondary, and tertiary response type; see Figure 2).

To manage the complexity of describing these findings, we group cases below based on their primary leaning (e.g., we refer to a reflective-compliant case in the reflective response section) and describe trends associated with the primary response type. In what follows, we describe the kinds of responses evident in more reflective, compliant, and distortive cases, within three dimensions central to the Compass policy: the perceived purpose of evaluation, observation procedures and support for teacher practice changes, and goal-setting practices. Table 2 summarizes the general patterns observed within the three overarching response types.

Two final notes are in order. First, we recognize that schools may engage in more reflective, compliant, and distortive behaviors generally, but this analysis focuses solely on how educators responded to the teacher evaluation policy. Thus, a case categorized as distortive in its response to evaluation, may in fact engage in other practices to facilitate teacher reflection and improvement, but these fall outside the scope of this research.[9]

*Distortive Responses.* All of our cases except one demonstrated some distortive responses to

TABLE 2

*Patterns of Response Type Across Case Characteristics*

| Case | Authorizer | Eval. system | Overall response type | Response characteristics | Leadership | Collaborative structures |
|---|---|---|---|---|---|---|
| Gull | OPSB charter | Waiver | Reflective | Commitment to evaluation and the goal of continuous growth | Shared, hands-on leadership (proactive support provision) | Frequent, purposeful collaboration for improvement |
| Nighthawk | RSD charter | Waiver | Reflective | Rigorous observations, support for growth | | |
| Plover | OPSB charter | Basic State Model | Reflective / Compliant | Meaningful personal goals for all educators/ students | | |
| Heron | RSD charter | Basic State Model | Compliant / Reflective | "I'm going to be effective anyway" | Few administrators, hands-off leadership (support available if needed), buffer teachers from policy | Casual collaboration and use of "experts" |
| Flamingo | OPSB direct-run | Basic State Model | Compliant / Reflective | "She tells you something negative, you make the correction" | | |
| Shearwater | RSD charter | Basic State Model | Compliant / Reflective Distortive | Irrelevant, standardized goals | | |
| Longspur | RSD charter | Basic State Model | Distortive / Compliant | Skepticism around the validity and purpose of evaluation | High autonomy, employee–manager relationship, buffer teachers from policy | High levels of autonomy, minimal collaboration |
| Finch | OPSB direct-run | Basic State Model | Distortive | Putting on a show<br>Playing a numbers game | | |

*Note.* OPSB = Orleans Parish School Board; RSD = Recovery School District.

evaluation policy, but none of our cases qualified as fully distortive in character. Two cases exhibited enough distortive responses to qualify as primarily distortive in character: Finch (distortive, with 56% of words coded for response identified as distortive across 60% of interviewed teachers and no interviewed administrators) and Longspur (distortive-compliant, with 16% of words coded for response identified as distortive across 100% of interviewed teachers and 67% of interviewed administrators). Both adopted the basic state evaluation model (i.e., five-domain rubric, two observations, no waivers). In both cases, the distortive responses precluded reflection by taking actions that led to invalid—real or perceived—measures of teaching practice.

*Skepticism around the validity and purpose of evaluation*. Overall, educators in distortive schools expressed skepticism about the purpose of Compass and the validity of its measures. One common reason for this skepticism was a belief that measures were not rigorous enough or adequately adapted for specific classrooms. Some also reported that it was too easy to "game the system." For instance, when asked how teachers respond to evaluation in the school, one administrator shared, "I think they're jumping through the hoops because they know they have to . . . It's not setting a bar that's making everybody want to rise to get there." Despite the stated intent of Compass to increase expectations around instruction, implementation at individual schools may not have set a high bar for teacher effectiveness or met expectations of rigor.

Some teachers (i.e., in special education or certain nontested subjects) at distortive cases reported that observation rubrics were not well suited to their instructional setting. For example, teachers who worked with students on a pull-out basis, often providing tailored support in makeshift classrooms, struggled to demonstrate rigor in lesson planning or classroom climate. One teacher shared, "I like being evaluated, but I just want a fair evaluation. That's all I'm asking . . . we're still trying to fit a square peg in a round hole." As a result, special education teachers at our distortive cases did not believe that the observation rubric provided a valid measure of their teaching quality.

What distinguished distortive schools was that teachers reported changing their behavior during evaluations to enhance their results. Rather than improving their practice to meet high expectations (a reflective response) or continuing their practice as usual (a compliant response), teachers felt compelled to engage in strategic behavior to appear effective according to the evaluation criteria (a distortive response). Described by some as a "dog and pony show," this behavior curtailed the opportunity for meaningful feedback and improvement. As one teacher stated,

> It's almost like a game; you got to learn how to play . . . I just say every time I do it, I get a little better, and not necessarily with my teaching practice but with what they're looking for.

We describe these behaviors further below.

*Putting on a show*. Some strategic behaviors were intended to ease the strain of evaluation on teachers and administrators alike. For example, at Longspur, administrators created a lesson plan template tailored to meet the Compass rubric standards. Administrators emphasized helping teachers get "very well prepared" for lesson planning and preparation for observations. Although coaching teachers in lesson planning can be a very helpful activity, these behaviors appeared to focus on creating a one-time event—the evaluative observation. A few educators noted that some teachers could put on a "dog and pony show" during observations to meet rubric criteria without demonstrating their typical practice. Similarly, a teacher reported selecting and adapting lessons better aligned with the Compass rubric on observation days to earn a higher score:

> Sometimes you may have a structure in your classroom that is working very, very well and students are learning from it, and it really fits the flow of the class, but for it to fit into the rubric, you might have to tweak it just for that particular lesson, which can be frustrating.

Teachers also reported other specific strategic behaviors to earn high scores on observation measures. In one example, a pull-out teacher being evaluated made arrangements to take over a regular teacher's class to demonstrate instruction of a planned lesson for the announced observation. Although this structure helped the administrator use the basic state rubric to evaluate the teacher, any feedback received would likely be irrelevant to the teacher's daily practice, which normally occurred outside of the regular classroom setting. In a dramatic example, a teacher at a distortive case school reported that she and others sometimes sent certain students to different classrooms during observations to avoid disruptions and ensure that lessons addressed all rubric criteria[10]:

> Sometimes as teachers, what we do is we will say, "Okay, we have a child that's not having such a good day, maybe you'll hold them while we're doing that [observation]." You have to do that honestly because you don't want to be penalized . . . We might have them go to another teacher and give them that assignment. Technically they couldn't handle the activity anyway. I know it's terrible, but it's reality. The children know; they know.

These strategic behaviors make valid measurements of teaching impossible and are unlikely to generate useful feedback from observers. Moreover, in these distortive cases, several teachers reported that they had not received feedback or coaching following their evaluative observation.

*"Playing a numbers game*." At distortive-leaning cases, teachers worked hard to ensure that they would meet their target SLTs, sometimes in strategic ways. According to administrators, teachers often set low SLTs to ensure that they would not be penalized for failing to meet rigorous targets. In some cases, teachers reported identifying students (and content) they felt they could best improve and focusing additional attention on just those students included in the SLT—a practice echoing the "bubble kids" phenomenon observed in classrooms of teachers responding to test-based, school-level accountability (Booher-Jennings, 2005; Hamilton et al., 2007). At some distortive schools, this meant that teachers selected a subgroup of students from just one of their class periods to which they provided extra support. One teacher explained, "Then I'd look at a focus of 30, maybe one class out of my three classes, half of the 30: 15 out of 30 should be able to move from basic to above [ratings on the assessment]." Teachers appreciated the sense of control that came from being able to focus their efforts and move student scores, but these kinds of decisions have the potential to lead to inequitable access to quality teaching for students not included in an SLT and to a skewed measurement of teachers' contribution to student success. One teacher provided a detailed example:

> It's almost like you're playing a number game [with SLTs]. Essentially, again, I'm teaching all of my kids, but which ones am I really focusing on to make sure that they get the content needed for learning, to be successful and for the next year? . . . Who's going to give the most bang for your buck? . . . Let me focus on the kids that really need the help. The ones I could really move . . . I hate to say it like that, that's awful but that's the reality.

Although these behaviors reportedly made evaluation more tolerable for teachers, reflection was not evident in these cases.

*Compliant Cases.* Three cases exhibited primarily compliant tendencies in their response to evaluation. Educators in these schools appeared to go through the motions, but not reflect or act to improve practice. Although educators may have questioned the validity of measures, they did not resist, game, or adjust their practice to improve their perceived effectiveness or preclude valid measurement (as they would with a distortive response). The two compliant-reflective cases, Heron (33% of words coded for compliant response over all interviewed teachers and administrators) and Pelican (41% of words coded for compliant response over all interviewed teachers and half of interviewed administrators), and one compliant-reflective-distortive case, Shearwater (29% of response coding was compliant across two thirds of interviewed staff), utilized the basic state model for evaluation.

*"I'm going to be effective anyway*." In compliant cases, faculty varied in their perceptions of the purpose and validity of evaluation. Some educators viewed the evaluation system as a corporate-style reform, intended for external communication rather than an internal tool to gauge and improve teacher quality. One Shearwater teacher commented,

> [P]ressure from the public [leads to] the response from our educational leaders . . . that "here is something that we're doing" that put the public mind at ease, that we're doing our job as educators. Most of the time, it's just something that is required of us.

Most respondents in these cases expressed either ennui or trepidation at the thought of evaluations. For example, one administrator believed that teacher views of the system varied from, “it’s just another program that we’ll have for a couple of years” to “it really doesn’t matter. I’m going to be effective anyway.” This administrator, however, believed that the evaluation had the potential to validly measure teacher quality and encourage improvement. She said, “For some of those teachers I have to constantly remind them of ‘this is growth for you, and once you grow, you don’t really chop down anymore.’ Right? The growth keeps going.”

The compliant perspective was evident both when Compass was implemented faithfully (as one Heron administrator stated, “Compass? Mm-mmm. No. I never thought about it. I just do it. If you say this is what you have to do, do it.”) and when implementation was weakened (as one teacher shared, “Regardless if it’s Compass or if it’s something else, I’m going to still do my best. When it gets to the point where I can’t do my best, then it’s time for me to move on without Compass telling me that I need to move on.”). Although teachers in compliant cases did not change their teaching in response to the evaluation, they believed that certain activities garnered higher rubric ratings, but were not appropriate for teaching specific content, as this teacher shared:

> One of the times that they came out, they didn’t feel like I had the kids engaged enough, as far as group activities . . . I was introducing the kids to . . . a state-run website where the kids can keep up all of their transcripts and records and everything. I had it all up on the board, and I was going through the steps to show them how to create the account, how to do this and that and the other. They [the evaluator] didn’t feel like that was engaging enough. It wasn’t meant to be a group activity, per se, the day that they came into my class. That’s what I had on the lesson plan, and that’s what I was teaching. They felt like I should’ve had more of a group thing going on. For that [activity], it’s more individual.

Educators in these cases also found the policy cumbersome and time-consuming. A Heron administrator reported that the complex system was in fact less responsive than the previous one:

> It’s a lot of paperwork that’s involved. I have to type all of that stuff into the system and make comments on all of it whereas before I used to just go in and observe the class and call the teacher and then say, “I need you to do so-and-so, so-and-so,” and boom. That was it.

Furthermore, she questioned the usefulness of the rubric and process in identifying effective teaching:

> Without that piece of paper, following that rubric, you know what a good teacher is. You walk into the classroom, you know the good teacher. I mean kids are responding. You’re walking through the room and you see the objectives on the board . . . We knew that already. This is just a gauge that the state wants us to have . . . I’m thinking we just do what they say.

As these excerpts demonstrate, educators in more compliant-oriented cases did not respond very strongly (either positively or negatively) to the evaluation reform. Rather, educators reported continuing with prior practice. One administrator stated, “I’d go out on a limb and say [for] 80 percent of teachers, [the Compass evaluation] doesn’t affect them at all because they know, they know, they do, they do and it’s just another part of the process.”

*“She tells you something negative, you make the correction*.” Several educators at compliant-oriented schools viewed formal observations as more contrived and far less meaningful than informal observations. In particular, educators struggled with the basic state rubric. Some teachers had not seen it (“Haven’t seen it. I know they’re checking things off, but I’m not knowing what is being checked off.”), whereas others questioned its validity. One principal believed the rubric may have captured instructional techniques rather than student learning:

> I had teachers score ones and twos [on the rubric], but their kids were learning . . . [their instructional delivery was] all over the place, but their kids at the end of the day understood the objective and were able to answer the essential question on what was being asked. I have some teachers that on the tool hit threes and fours [on the rubric] because they’re very taskmaster: “My objective is posted; this is done; I’m going to transition well,” but content-wise, they’re not giving it to the kids at the level and differentiated instruction that they needed and so their kids are looking like, “What in the world are you talking about?” There’s no connection or there’s no learning.

To manage this disagreement, some administrators strictly followed the rubric but then softened the communication of ratings. One principal explained that "[when a] teacher's like, 'I'm a two.' Well, that doesn't mean you're a bad teacher. At this time, that's what we saw or that's what we felt you were. We just try to ease the tension about it."

Teachers at compliant cases also reported relatively minor modifications in response to ratings. One teacher at Heron reported that, "She [the principal] tells you something negative, you make the correction." A teacher at Shearwater reported that her principal's feedback was "just more rigor—that's the buzz word." As these responses indicate, little reflection took place as a result of observation feedback. Teachers at compliant schools may have made changes to their practice, but such changes generally were not made in response to evaluation measures or feedback. As one teacher shared, "Oh, I make changes all the time based on the evaluations or not."

*Irrelevant, standardized goals*. Compliant-oriented cases typically used SLTs as the only individualized goals recorded for each teacher and, in one case, used the same standardized goals across the school. One teacher did not agree with the standardized goal, but was not strongly affected as she maintained higher personal goals. She said,

> My objective every year is always to exceed that. I don't believe in mediocrity in any kid . . . They gave it [the SLT] to me. I have no idea where it came from . . . It's not going to affect me one way or the other.

In other cases, administrators helped teachers set "realistic" goals. One principal said,

> Sometimes we have to lower it [the SLT] because the anticipation, the expectations are just so high that . . . based upon what it is that we see with the test that we're picking that they may not make it.

Although intended to buffer teachers from a potential "ineffective" rating due to failing to meet high standards, these adjustments may have precluded authentic reflection. One teacher surmised that such a standardized system might even "drive away" effective teachers:

> You have to fill out these SLTs. You have to put them in. You have to have them approved. You have to use all the jargon. They'll say a lot of this paperwork is driving teachers away, even the best teachers.

*Reflective Responses.* Three schools responded to evaluation in more reflective ways. Educators in these schools reported engaging in meaningful reflection and improvement efforts, and clearly perceived the evaluation data as a valid measure of teaching and useful for improvement. While they reported SLTs and overall teacher ratings to the LDOE, two utilized enhanced evaluation procedures that supplemented the Compass requirements. The system design at Gull (reflective, with 77% of words coded at reflective across all interviewed staff) included more observations, the use of the full Danielson rubric, and extensive individualized coaching from mentor and master teachers. At Nighthawk (reflective, with 76% of words coded at reflective across all interviewed teachers and half interviewed administrators), the system also included more observations, replaced the state rubric with a set of position-specific categories of professional and instructional practice, and involved teachers in developing individualized growth plans and participating in frequent coaching from administrators. Plover (reflective-compliant, with 56% of words coded at reflective across all interviewed staff) utilized the basic state model for evaluation.

*Commitment to evaluation and the goal of continuous growth*. At reflective cases, educators reported strong levels of commitment to the validity and utility of evaluation processes. First, they believed their evaluation process and resulting data were accurate measures of quality. At Gull, one teacher described trusting, understanding, and believing in the value of their process:

> We do have inter-reliability . . . I think it's valid because we have rubrics that we are given based on lesson, structure within our lesson plans, culture. These rubrics and things are actually gone over with us several times throughout the year. Every teacher in this building has a mentor . . . We understand the rubric, we understand the expectations, and I think collectively, which makes us a dynamite school as well.

Furthermore, teachers described their commitment to the process as an important opportunity to gather feedback on instructional practice. "I know that these people are there," shared one teacher, "and that they're going to be honest, helpful, forthright and consistent with what they're scoring and what they're seeing." This teacher reported that feedback was given in good faith and intended to encourage improvement. Educators at these schools similarly reported that the purpose of evaluation was continuous improvement. One administrator at Nighthawk stated, "It's continuously growth. That's what we're looking for." A Gull teacher explained, "I mean who wouldn't want positive feedback on that? You're always going have something that you can grow."

*Rigorous observations, support for growth*. The kind of reflection described above is predicated on the expectation that failure is acceptable and that improvement is always possible and necessary for teachers to best serve students. As one Nighthawk administrator shared,

> We want to get better and better at it over time, but the goal is not to set these impossible benchmarks and then say, "Why aren't you doing this?" The goal is to say, "Okay, we have to get here. We know that's going to be really hard, so we have to build these things out over time in order to get there."

In practice, this acceptance of failure is bolstered by a sense of strong support. One teacher noted,

> [During observations] they're looking to support me and to see what things I'm doing well and all of that. I can look forward to the feedback that I'm going to get being something that's useful and productive, and not just, "Oops, that was your formal observation and you got all ones," or something really negative. It's a positive experience.

In part, teachers may have been motivated to improve because of the strong connection drawn between teacher practice and student achievement. As one teacher observed, "Yeah, if I don't make a mark on one of them [rubric elements], then I work on it. I try and get the help that I need to change it." Educators at reflective cases embraced the evaluation process as a meaningful and integral part of their professional practice and a useful tool for meeting student needs. One teacher shared, "We've seen really, how our teachers . . . become more and more reflective, and how that reflective process has really translated into more and more engaging lessons for the students."

In reflective cases, evaluators typically provided feedback by suggesting strategies and either modeling the intervention in the teacher's classroom or sending the teacher to observe a peer. These kinds of observation "fieldtrips" were tailored to the specific intervention at hand (e.g., the teacher would observe a peer who employed a particular strategy very well). At Gull, teacher leaders continually searched for innovative practice to address their teachers' shortcomings and piloted new practices to determine the fit with their community, culture, and style. As one teacher leader noted,

> We do all of the researching part of it, of what are best practices for the teachers. We field test it first before we bring it to teachers and doing this weekly professional development. We come back and model it to them, and we also offer them the opportunity, they can either ask us to come and model it for them, co-teach with them, or just observe and provide feedback to them.

Underlying these coaching strategies was the assumption that "best practice" may or may not be useful and appropriate for all teachers. In the reflective-oriented cases, evaluators emphasized that teachers should try out suggested strategies, but should also feel free to use alternative ones, and when one idea did not work, they worked with teachers to determine new strategies. For instance, one administrator described the improvement process as iterative and open to failure: "We give suggestions, and then expect the teacher to act on that. Sometimes acting on it works out. Sometimes not. If it doesn't, then we do the process again, and we go on from there." This practice of trial and error helped to communicate the culture of continuous improvement, and foster a sense of responsibility and empowerment among teachers to refine their practice.

*Setting and monitoring meaningful personal goals for all educators and students*. In reflective cases, teachers took goal setting seriously and differentiated goals for students. One administrator explained,

> Our teachers are very, very mindful . . . Let's say if they say [in their SLT that] 10 out of 15 students will meet a composite score of X number. Then what about the other five students? That second SLT goal focuses on those five students and what kind of growth are we going to have for these five students. All of the 15 students are being tracked, and they're being focused on intensely in there.

In these schools, educators regularly tracked goals and reflected on their progress. At Gull, teachers reflected weekly in journals about their students' progress on assessments and how they were going to improve on results. Such formative self-reflection was paired with self-reflection during the formal evaluation process. At Nighthawk, teachers rated their own performance on each rubric category prior to their evaluation conference and discussed any divergence from the administrator's scores. Together, goal setting and progress monitoring fostered reflection among case educators.

*Summary.* As discussed above, we find that schools in our sample displayed one or more response characteristics, ranging from distortive to compliant and, in three cases, reflective. From this initial look, it appears that local modification of the Compass policy was strongly related to reflective responses. As noted, two of the three reflective case studies utilized locally selected methods of evaluation that extended beyond the basic state model. These customized procedures may have engendered increased buy-in to the evaluation system among educators and promoted reflective responses. Of course, the directionality of this relationship is unclear. It is equally possible that reflective schools were more motivated to invest in customized evaluation procedures. In contrast, the compliant and distortive cases all utilized the basic state rubric, perhaps due to ambivalence toward the state mandate. Beyond this one distinction, it is not clear that any one type of local adaptation or design feature led to more reflective response. In what follows, we examine how other organizational factors influenced responses to teacher evaluation.

### *Organizational Factors Related to Variable Responses*

At the outset, we expected to see differences in the implementation of Compass and local evaluation systems according to school authorizer and governance model, as these characteristics were thought to influence school-level autonomy, history, and other contextual factors. For example, we expected to see different kinds of evaluation responses in single-site RSD-authorized charters compared with OPSB direct-run schools or charter schools in larger CMOs, due to their varying capacity and level of autonomy. Schools supported by a CMO or OPSB might have had more administrators available to support evaluation implementation, whereas single-site charter schools might have had greater flexibility to design a local evaluation system and waive the basic state model. Our matrix analysis, however, revealed no clear patterns (see Table 2). In a set of secondary analyses, we also looked for response patterns according to basic school characteristics such as size, level, performance, and demographics, as well as individual respondent characteristics, including gender, race, certification type (traditional or alternative), and years of experience (overall and at the school). We found no clear pattern of responses by any of these categories.

We did, however, see clear patterns of variation in school response related to two school-level organizational factors. We found that *quality of leadership* and *structures for collaboration* were strongly related to the response type of the cases, whereas there was no clear pattern of response according to school history with evaluation, level of autonomy, or school culture. As such, we focus on the aspects of leadership and collaborative structures that potentially contributed to, and/or resulted from, different responses to evaluation (see Table 2).

*Leadership.* Our matrix analysis illustrated a strong relationship between shared and hands-on instructional leadership and reflective responses. In our reflective cases, school leadership appeared to enhance capacity to complete a meaningful evaluation process and accountability for continuous improvement. In practical terms, shared leadership—the inclusion of additional administrators and teacher leaders in the management of instruction and operations—enabled schools to complete teacher evaluation in reflective ways by expanding the number of evaluators and support providers. Most notably, Gull provided teachers with a tiered career pathway where they could begin to take on more mentorship and evaluation responsibilities. As such, multiple teacher leaders had the time and motivation to evaluate and support mentees. One administrator noted that shared leadership provided the flexibility to tailor support to individual teachers:

> If it's just somebody who really needs additional support, then it's probably going to be that master teacher who can really have the time to get in the classes and really monitor what's going on, and offer that advice. Then it also may be a mentor who either works on the same grade level as that person, who can really offer insight and support, or it maybe someone who is not on that grade level, but has a vast knowledge of ELA, math, who can really help on that curriculum aspect . . . [I]t could just be relationships. If I have a really great relationship, where I know I can be very honest, and give you some feedback that may not be very palatable, you're going to take that a little better coming from a peer, or coming from someone who you understand.

In these schools, this increased capacity—due to leadership sharing among administrators and teacher leaders—granted each evaluator enough time to thoughtfully complete rubric ratings and provide support. At Gull, evaluators reported spending time after each observation carefully rating each teacher based on the evidence at hand—rather than assigning ratings during the observation, which was more commonly reported in nonreflective schools.

Furthermore, hands-on leadership—a construct that emerged during our analysis describing frequent communication between administrators and teachers regarding teacher practice—in reflective schools encouraged evaluators to spend time purposefully planning for their meetings with teachers individually and as a group. At Nighthawk, evaluators planned for each weekly conference with teachers and sent out meeting agendas in advance. As one administrator noted, "We have the [teacher's individual] goals at the top of our meeting notes and then as we break things down over time, we think about, are we actually making progress towards these goals?" This kind of purposeful planning and clear communication on the part of leaders lent structure to the meetings and also communicated a sense of accountability. Teachers and administrators reported being constantly aware of their progress toward goals and the implementation of planned interventions. Although this leadership style might appear intense, perhaps verging on "micro-management," case study teachers appreciated this approach and heavy administrator involvement. In fact, at Plover, one teacher described the school leader as "notorious" for being "very assertive in a powerfully good way."

In part, this hands-on, directive leadership seemed to work because of what Elmore (2002) called "reciprocity of accountability for capacity." That is, teachers understood that while they were responsible for improvement, school leaders were equally responsible for helping build their capacity to improve. The schools' commitment to coaching provided this assurance. Evaluators took on the role of coaches to teachers, using processes widely cited in the literature, including setting goals, suggesting high leverage practices, describing, modeling, and reflecting on the use of practice (Knight & van Nieuwerburgh, 2012; Marsh, Bertrand, & Huguet, 2015; Marsh et al., 2008). As one teacher leader at Gull stated,

> We want to foster that reflective process. Then that's where I would go through my evidence of this is what I saw . . . that was really, really done well or really good. Please continue to do this . . . This is something that I think could be improved. Here is the evidence for the statement of why I say it could be improved. Or here is something that could, if just adjusted, could impact student achievement even more . . . Sometimes that may be used as a refinement area for work on, may not be something that they score a two on. It could be a three, and it could be a four. It's just in this particular lesson, if you just tweak this a little bit, it could further impact student achievement more.

In this example, the evaluator discussed using questioning to facilitate teacher reflection, before suggesting strategies to affect student achievement. Notably, even the numerical ratings were used as yard lines (how far along the field you moved the ball), rather than goal posts (scoring or not).

Similarly, a Nighthawk administrator described how evaluators chose the highest leverage strategies to focus on, while acknowledging the consistent need for improvement across all areas:

> What we try to say is, you're going to have twos and threes and fours in lots of different areas, but we don't

> want you to work on all those things at the same time. We're going to focus on this particular area. That doesn't mean that you don't also have to do those other things . . . Usually, if you do that, people then tend to up their practice and the other things tend to rise as you put areas of focus on the things that seem concerning.

After each observation, evaluator-coaches at these reflective cases set short-term goals, rather than longer term annual goals, and identified a limited number of practices to improve. Rather than identifying a deficit and asking teachers to find a solution, evaluators suggested immediate ways to improve practice. One teacher explained,

> Because our master teachers will go out of their way to say, "Oh look, I saw this strategy here, I wanted to see if you want to use it or not." Or, "This is what I used to do in my classroom. We used to use this, and it worked really well in small groups" or stuff like that . . . It's not like you're just thrown in a group and you're like, "Do what you need to do."

In contrast, across the compliant response cases, administrators retained the bulk of the leadership responsibilities with support from a small number of support administrators and/or teacher leaders. Whereas compliant case administrators exhibited a range of leadership styles, the principals at three of the cases generally provided hands-off leadership (in their words, "let people be them") while remaining available for support and advice. In a sense, these principals served as experts on call. One teacher at a compliant case shared, "There's always an open door policy, so if I ever have any problems I can go to either one of them at any time, if they're available."

One challenge associated with the limited shared leadership at these sites was having adequate capacity to complete meaningful evaluations and provide individualized support. Administrators at all compliance-oriented sites acknowledged having insufficient time to complete evaluation observations, monitor teacher improvement, and provide individualized support to teachers. When asked about evaluation work such as observations and coaching, one teacher recognized this challenge, saying,

> I think if we had more staff, more teachers, then I think the workload would be more manageable. I think people would be happier, and I think that learning would take place in a—at a higher rate, and things would be better.

In our distortive cases, by contrast, teachers often appreciated their autonomy and embraced an employee–manager relationship with administrators. In these cases, teachers stated that administrators did not "micromanage," rather letting teachers "do what you need to do . . . then, if you don't, then she tightens up." Said another way, one teacher at a distortive case said of the administrator, "If you're doing your job, you're doing what you're supposed to, [the principal is] very easy to get along with." One teacher reported wishing for additional support, reciprocity of accountability, and a sense of shared purpose:

> I think a lot of people feel this. If we felt like all of the people above us who are, right now, it feels like are giving us these directions to do—if we felt like they also felt accountable for our growth and for our students' growth, then it would—that's another piece that would really make it feel cohesive, like we are a team together because not only—I want to be accountable. I want someone checking in and helping to remind me, "Oh, remember, do this." "Oh, yeah-yeah," because I need that support. I want to succeed, and I want my kids to succeed, but I also want to know that you're doing the same thing.

Much in the same vein, administrators in these sites focused on reviewing data with teachers, rather than providing instructional coaching. Limited time challenged the ability of principals to provide detailed, personalized feedback and support to teachers. Furthermore, the hands-off style of leadership likely did not create a culture conducive to experimentation, failure, and reflection.

In both compliant and distortive cases, leaders also took on the role of buffering teachers from the evaluation policy. In compliant examples, these behaviors included administrators counseling teachers that any score in the "effective" range (3 or 4) is perfectly fine. While attempting to allay teacher fears regarding the consequences of observation scores and encourage teachers to focus less on "the grade," this strategy might have inadvertently discouraged teachers from engaging in reflection and seeking support to improve their practice. In distortive examples, buffering was evident as administrators counseled teachers to purposely set low student achievement goals, to guard against the possibility of an ineffective rating because, “we know that there are some very dangerous consequences for failure that are beyond our control.” Thus, teachers were unable to use their evaluation ratings to reflect on their effectiveness and growth.

*Structured Collaboration.* Structured collaboration (i.e., consistent, purposeful time reserved for teacher-to-teacher interaction) also emerged as an important factor shaping evaluation in case schools, working hand in hand with shared leadership. In our cases, professional collaboration was organized to allow for three distinct types of interaction: (a) frequent, purposeful collaboration for improvement (associated with reflective cases); (b) casual collaboration and use of “experts” (compliant cases); and (c) minimal collaboration (distortive).

Our reflective cases scheduled purposeful and consistent time for teachers to meet, assigned teachers or administrators to facilitate collaborative discussions, developed tools to aid discussions, and communicated expectations that teachers engage in such practices regularly. As noted, instead of merely asking teachers to meet during a specific time period, school leaders arrived at teacher collaboration meetings with an agenda, guiding questions, and procedures for examining student work and data, reflecting on practice, and brainstorming solutions. Below, one teacher described how collaborative meetings enabled teachers to share strategies across grade levels, search for external resources, and to observe or have lessons modeled:

> Those weekly meetings that come in, there’s so much collaboration that goes on there . . . One of the mentor teachers will demonstrate or model, whatever’s expected, or whatever we might be struggling with . . . They’ll pull up information off of the internet . . . If you need me to come in and do a lesson for you, we schedule it out that week . . . I take whatever I need to bring myself up to the level I need to be on.

As this final sentence demonstrates, this kind of collaboration appeared to foster a sense of accountability for continuous improvement. Collaborative meetings also allowed teachers to monitor their progress in using a new practice in the classroom. According to one teacher leader, “We have to start them off at this level with the strategy, and then watch them grow with it, and then record their progress. Each week we come back with that information,” and when more progress is needed, “I need to sit down and work with them a little bit more on this strategy. It’s very beneficial.” Opportunities for collaboration also provided an avenue for enhanced peer accountability.

In the more compliant cases, most educators reported that colleagues in their schools had a “shared vision” and viewed their work as a “team effort.” Furthermore, teachers reported that they could access support, help, materials, and ideas from fellow teachers. Yet, on the whole, collaboration was not purposive and there were few structures in place to facilitate regular collaboration. Instead, teacher interaction occurred informally and peer support was provided only to those who asked for it. One teacher noted, “There could be more collaboration, more time allotted for teachers to work together.” At one compliant case, the only structured avenue for sharing practice was through the identification and use of “in-house expert” teachers. Because these “experts” were identified by administrators alone and only permitted to assist with particular topics, the flow of knowledge and support was severely constrained. In fact, the selection of “experts” might have implied to teachers that they should only collaborate with identified individuals upon an administrator referral. Without structured, frequent opportunities to meet, teachers had limited access to learn from others.

In more distortive cases, communication among educators was at times strained, and collaboration was informal at best. “Things like structures for communication and decision making among administration and staff has become difficult,” said a teacher from one distortive-leaning school, “I find that as something that we’re still working on.” Although collaboration around student behavioral issues and new instructional strategies occurred in some departments, in other departments, teachers had minimal contact with their peers. In these cases, teachers missed opportunities to engage with colleagues for reflection and growth. Across our cases, there were strong indications that, together, evaluation design, leadership, and collaborative structures contributed greatly to the types of responses to evaluation exhibited by schools.

### *Cross-Cutting Patterns: School Goal Orientations*

Looking at the practices and associated school-level contextual factors across cases, we are struck by an overarching set of differences in school-level goal orientations that relate to our response typology. Goal orientations are "reasons and purposes for approaching and engaging in achievement tasks" (Pintrich, 2003, p. 676). Contrasting mastery and performance goal orientations in students, Schraw (1998) summarizes, "Students with mastery orientation seek to improve their competence. Those with performance orientations seek to prove their competence" (p. 122). A mastery goal orientation focuses individuals' attention on "developing new skills, trying to understand their work, improving their level of competence, or achieving a sense of mastery based on self-referenced standards," whereas a performance goal orientation focuses attention on achievement relative to others (Ames, 1992, p. 262). Typically applied to students, here we see relevant applications of the concept of goal orientations to educators.

Whereas past research identifies classroom structures that promote such orientations in students (Ames, 1992; Epstein, 1988; Urdan & Schoenfelder, 2006), our exploratory research suggests that schools as a whole may also foster these orientations in ways that affect teachers and leaders. Overall, our reflective cases demonstrated a set of practices and an organizational culture that promoted and valued a mastery goal orientation for teachers, whereas the compliant and distortive cases reflected a performance orientation. In our reflective cases, we saw a strong focus on continuous improvement and innovation, facilitated by a comfort with failure. Implicit in these schools was an understanding that with effort, all teachers could achieve at high standards. This mastery orientation was evident in the evaluation tools and practices encouraging teachers to set short-term, self-referenced goals and to take responsibility for goal setting. As such, teachers at reflective cases did not focus on reaching a particular rubric rating, but instead on honing their practice and experimenting with new ideas.

In contrast, our compliant and distortive-leaning cases demonstrated a performance goal orientation for teachers. This was evident in how teachers and administrators conceptualized their own targets and capacity, as well as those of students. For example, setting of SLTs by administrators essentially discouraged teachers from taking responsibility for the goal-setting process and setting ambitious targets based on their own improvement. Rather, such a system incentivized teachers to simply meet a moderate threshold of student performance. Student performance was often described as being outside of the teacher's control, linked to static student characteristics. Similarly, some teachers exhibited a fixed perspective on their own capacity, one stating that "I can't do grouping," and accepting a lowered observation rating rather than trying to learn skills for instructional differentiation.

In schools with performance-oriented environments, we also observed evidence of goal displacement—an outcome commonly found in organizations responding to accountability and reward systems (Hentschke & Wohlstetter, 2004; Kerr, 1975). Strategic behaviors—such as sending "difficult" students out of the classroom during an observation, setting goals based on the performance of only students expected to show growth, and designing lessons specifically to meet rubric criteria—evidenced a displacement of the goal from that of improving teacher practice or student achievement, to that of scoring an effective evaluation rating. Actions were designed to reach a high score ("getting a four"), rather than to improve instruction or student achievement. Indeed, Longspur administrators described teachers changing their practice with words like "they nailed the rubric," rather than "they improved their teaching." This kind of goal displacement, particularly as teachers focused on addressing just five Compass rubric elements at the possible expense of general improvement of teaching, parallels the narrowing of curriculum and instruction in response to high-stakes testing accountability regimes (e.g., Darling-Hammond & Wise, 1985; Hamilton et al., 2007). Of course, given the limited scope of our study, we cannot rule out the possibility that everyday instruction encompassed a broader range of practices than the rubric lists.

## Conclusions and Implications for Policy

To summarize, our case studies demonstrated a range of responses to evaluation policy, from reflective to compliant to distortive (Research

Question 1). We found that these responses were strongly related to aspects of leadership and collaborative structures. Specifically, hands-on instructional leadership, shared among administrators and teacher leaders, along with frequent, structured teacher collaboration, was associated with reflective responses to a state-mandated teacher evaluation policy (Research Question 2). Reflective cases also appeared to have a more mastery-oriented environment. In addition, we found that schools that modified the state-recommended evaluation design were often more reflective in their responses. The direction of all of these relationships, however, is unclear.

Together, our findings give rise to three sets of implications regarding the implementation of teacher evaluation policy in and outside of New Orleans. The finding that only three of eight case study schools engaged in primarily reflective practice through Compass suggests that significant effort is needed to elicit meaningful teacher improvement through MMTES. In presenting these ideas, we recognize that Louisiana policymakers are limited in their ability to shape implementation of state policy in the highly decentralized system of schools in New Orleans and that policy options likely differ in more "typical" state and district settings. Nevertheless, the broad ideas are still relevant to all education leaders interested in using evaluation for accountability and improvement.

First, this study suggests the need for policies and resources that foster organizational conditions associated with reflective responses. As our findings indicated, shared leadership and structured collaboration appeared to promote greater learning and mitigate the burden on administrators to observe, evaluate, provide feedback to, and support teachers. Capitalizing on its significant autonomy, New Orleans schools could further develop and innovate models of leadership and collaboration that promote organizational and individual learning tied to observation and SLT data. Local leaders might also consider ways to allocate resources to teacher leader positions, shared planning time, and tools that foster collaborative discussion tied to evaluation results. In settings beyond New Orleans, where teachers are often prohibited by collective bargaining agreements from conducting formal evaluations, education leaders should consider allowing more experienced teachers to coach their peers to learn and improve based on evaluation results. Policymakers might also adopt policies and allocate resources that allow for distributed leadership, teacher collaboration, coaching, and career-ladder programs. Coupling MMTES with these supports might also build much-needed teacher buy-in by guaranteeing reciprocity in the accountability arrangements of MMTES. That is, teachers may be more receptive to mandated evaluation standards if they are assured that they will receive the support needed to achieve them.

Second, our research suggests that flexibility to modify evaluation policy may promote greater organizational learning. As noted, schools that adapted the state model, using more detailed, expansive rubrics and added observations, tended to exhibit more reflective responses. Although we cannot prove causality, the fact that state policy allowed for this flexibility certainly provided opportunities for customization and greater buy-in. This finding may affirm the direction set by the new ESSA policy's emphasis on local control. By not mandating overly prescriptive policies and allowing policymakers to design their own educator evaluation systems with a commonly established framework for rigor, states and LEAs may see enhanced teacher buy-in and implementation.

Finally, our study suggests policymakers consider potential tradeoffs as they design and revise the elements within MMTES. One important choice in the Compass reform was to include only five Danielson elements in the observation rubric to ease the burden on administrators. This decision may have limited the comprehensive picture of the quality of teaching, as the selected standards focused primarily on measuring instructional quality according to student behaviors, and may have encouraged strategic behavior (e.g., gaming) and precluded reflection around other elements of teaching. To mitigate this issue, policymakers may consider alternative approaches to using a comprehensive rubric without placing undue burden on school personnel, such as rotating the rubric elements assessed each year. Policymakers may also consider providing a menu of possible rubrics that schools can tailor to their needs and preferences, or allowing a mix of required and teacher-selected rubric elements for inclusion. It behooves policymakers to keep these tradeoffs and potential limitations in mind when selecting elements to include in MMTES.

### *Directions for Future Research*

This study was exploratory in nature and limited by the small number of case schools and use of cross-sectional case study data. Future research might build on this work in important ways. First, longitudinal data collection would enhance our understanding of organizational learning. In particular, it may be interesting to examine changes in implementation and organizational context over time. Could rigorous teacher evaluation policies eventually modify organizational context and/or goal orientations? One might hypothesize that the experience of "putting on a show" could serve as a mastery experience, facilitating teacher learning. As such, even distortive responses might lead to organizational and individual learning, albeit at a slower pace. Future studies might examine how these themes play out on a wider scale and over a longer time period. Just as scholars have begun to identify classroom structures promoting mastery, there also may be opportunities to identify school-level structures that facilitate mastery for teachers. Scholars might also consider incorporating observations of school and teacher practice into future studies to gain a deeper understanding of the mechanisms promoting various goal orientations for teachers and the organizational routines that contribute to more versus less reflective responses to teacher evaluation.

Our study also leaves unanswered questions about the substantive effects of school responses to evaluation policy on schools, teachers, and students. Future studies might examine the extent to which distortive, compliant, and reflective responses lead to organizational change, new teacher practices, and ultimately improved outcomes for students. Whereas theory would predict more superficial or a lack of effects within distortive and compliant cases, empirical studies could document the nature, magnitude, and sustainability of effects across different types of schools. Comparative studies of schools within other districts would also be beneficial. The distinction between charter and traditional schools may be less pronounced in New Orleans given that all schools operate in a highly competitive context. Future studies might examine organizational responses to evaluation in other cities with more versus less decentralized management.

## Acknowledgments

We thank Douglas Harris, Alica Gerry, Nathan Barrett, and other Education Research Alliance (ERA) New Orleans colleagues for assistance with this research. Institutional support was provided by Tulane University's Murphy Center and School of Liberal Arts.

## Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was conducted at the Education Research Alliance (ERA) for New Orleans at Tulane University with support from the Laura and John Arnold Foundation and William T. Grant Foundation.

## Notes

1. Here we draw on Rondinelli's (1981, 1989) framework of multiple types of decentralization. Deconcentration is the transfer of authority from a central government agency to a local branch of the central government. Delegation transfers authority from one public agency to another. Devolution meaningfully transfers authority to a lower governmental unit (e.g., from the state to a local school district). Privatization gives authority to nongovernmental agencies such as charter management organizations (CMOs). Throughout this article, we use the generic term *decentralization* to refer to the latter two types. Specifically, we define decentralization as state or local education agency giving authority to school managers whether at traditional government-run public schools (devolution) or privately run charter schools (privatization).

2. Created in 2003, the Recovery School District (RSD) was empowered to take over and turn around low-performing schools in Louisiana. After Hurricane Katrina in 2005, the RSD became a major vehicle for restructuring the New Orleans school system.

3. Author calculations from student and teacher data provided by the Louisiana Department of Education.

4. In 2016, Louisiana's new governor, John Bel Edwards, signed into law a new policy that will return all schools overseen by the state RSD (not Louisiana Board of Elementary and Secondary Education [BESE]) to the local Orleans Parish School Board by 2018–2019. These will remain charter schools, but under local oversight.

5. Most of those used a more comprehensive approach by applying the full Danielson Framework or the Teacher Advancement Program (TAP) program (Danielson, 2013; Glazerman & Seifullah, 2012).

6. We chose to call this category "reflective" instead of "productive," as Jennings (2012) does, because it more precisely describes interactions that promote individual and organizational learning (e.g., Schön, 1987). For example, from the perspective of a school leader, it might be productive to engage in a distortive response if it advances school goals.

7. Extensive research indicates that individual capacity (skills, knowledge, dispositions, sense of self, etc.) may influence responses to and implementation of policy (Bandura, 1977; Beaver & Weinbaum, 2012; Massell, 1998; O'Day, Goertz, & Floden, 1995; Spillane & Thompson, 1997; Stoll, 1999, 2009). We include individual capacity in our theoretical framework as an important dimension of the theory presented; we do not, however, measure these elements of individual capacity due to resource constraints.

8. We are unable to provide details on the characteristics of individuals interviewed at each case school, as this would compromise anonymity. In Table 1, we provide aggregate characteristics.

9. Our analysis is not intended to be normative or evaluative. Instead, we anchor the analysis in the perspective of state policymakers, as communicated in Bulletin 130, and examine to what extent school implementation aligned with the explicit goals of Compass, which by design embraced reflective intent. Although one could argue that a distortive or compliant implementation of Compass might be "good" if one determines Compass is a "bad" policy, such an interpretation represents a different approach to the one taken herein.

10. To further protect the anonymity of respondents, we use the pronouns *she/her* to respond to all interview respondents regardless of their gender and pseudonyms to identify our case schools.

## References

Act No. 54, House Bill No. 1033, Louisiana House of Representatives. (2010).

Advisory Committee on Educator Evaluation. (2011). *Compass: Louisiana's path to excellence*. Baton Rouge: Louisiana Department of Education, Office of Innovation, Advisory Committee on Educator Evaluation.

Amabile, T. M. (1997). Motivating creativity in organizations: On doing what you love and loving what you do. *California Management Review*, *40*, 39–58.

Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology*, *84*, 261–271.

Arce-Trigatti, P., Harris, D., Jabbar, H., & Lincove, J. (2015). Many options in New Orleans choice system. *Education Next*, *15*(4), 25–33.

Baker, A. (2013, December 22). Bumpy start for teacher evaluation program in New York Schools. *The New York Times*. Retrieved from http://www.nytimes.com/2013/12/23/nyregion/bumpy-start-for-teacher-evaluation-program-in-new-york-schools.html

Bandura, A. (1977). *Social learning theory*. Englewood Cliffs, NJ: Prentice Hall.

Beaver, J. K., & Weinbaum, E. H. (2012). *Measuring school capacity, maximizing school improvement* (CPRE Policy Brief No. RB-53). Philadelphia, PA: Consortium for Policy Research in Education.

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment*, *17*, 62–87. doi:10.1080/10627197.2012.715014

Blase, J., & Blase, J. (1999). Principals' instructional leadership and teacher development: Teachers' perspectives. *Educational Administration Quarterly*, *35*, 349–378.

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

Burian-Fitzgerald, M., Luekens, M. T., & Strizek, G. A. (2004). Less red tape or more green teachers: Charter school autonomy and teacher qualifications. In K. E. Bulkley & P. Wohlstetter (Ed.), *Taking account of charter schools: What's happened and what's next?* (pp. 11–31). New York, NY: Teachers College Press.

Campbell, J. L., Quincy, C., Osserman, J., & Pedersen, O. K. (2013). Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological Methods & Research*, *42*, 294–320.

Chen, C. J., & Huang, J. W. (2007). How organizational climate and structure affect knowledge management—The social interaction perspective. *International Journal of Information Management*, *27*, 104–118.

Chetty, R., Friedman, J., & Rockoff, J. (2014a). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, *104*, 2593–2632.

Chetty, R., Friedman, J., & Rockoff, J. (2014b). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review*, *104*, 2633–2679.

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

Coburn, C. E. (2004). Beyond decoupling: Rethinking the relationship between the institutional environment and the classroom. *Sociology of Education*, *77*, 211–244.

Cravens, X. C., Goldring, E., & Penaloza, R. (2012). Leadership practice in the context of U.S. school choice reform. *Leadership and Policy in Schools*, *11*, 452–476.

Cuban, L. (1988). *The managerial imperative and the practice of leadership in schools*. Albany: State University of New York Press.

Danielson, C. (2013). *The framework for teaching evaluation instrument, 2013 edition: The newest rubric enhancing the links to the Common Core State Standards, with clarity of language for ease of use and scoring* (2nd ed.). Princeton, NJ: Charlotte Danielson Group. Retrieved from https://www.danielsongroup.org/framework/

Darling-Hammond, L., & Wise, A. E. (1985). Beyond standardization: State standards and school improvement. *The Elementary School Journal*, *85*, 315–336.

Dee, T. S., & Wyckoff, J. (2013, October). *Incentives, selection, and teacher performance: Evidence from IMPACT* (NBER Working Paper No. 19529). Cambridge, MA: National Bureau of Economic Research.

Derrington, M. L. (2014). Teacher evaluation initial policy implementation: Superintendent and principal perceptions. *Planning and Changing*, *45*, 120–137.

Derrington, M. L., & Campbell, J. W. (2015). Implementing new teacher evaluation systems: Principals' concerns and supervisor support. *Journal of Educational Change*, *16*, 305–326.

Dodgson, M. (1993). Organizational learning: A review of some literatures. *Organization Studies*, *14*, 375–394.

Doherty, K. M., & Jacobs, S. (2015). *State of the States 2015: Evaluating teaching, leading and learning*. National Council on Teacher Quality. Retrieved from http://www.nctq.org/dmsView/StateofStates2015

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

Donaldson, M. L., Cobb, C. D., LeChasseur, K., Gabriel, R., Gonzales, R., Woulfin, S., & Makuch, A. (2014). *An evaluation of the pilot implementation of Connecticut's System for Education Evaluation and Development*. Neag School of Education. Retrieved from http://www.connecticutseed.org/wp-content/uploads/2014/01/Neag_Final_SEED_Report_1-1-2014.pdf

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

Doyle, M. C., & Feldman, J. (2006). Student voice and school choice in the Boston pilot high schools. *Educational Policy*, *20*, 367–398.

Ellett, C. D., & Teddlie, C. (2003). Teacher evaluation, teacher effectiveness and school effectiveness: Perspectives from the USA. *Journal of Personnel Evaluation in Education*, *17*, 101–128.

Elmore, R. F. (2002). *Bridging the gap between standards and achievement*. Washington, DC: Albert Shanker Institute.

El Sawy, O. A., Gomes, G. M., & Gonzalez, M. V. (1986). Preserving institutional memory: The management of history as an organizational resource. *Academy of Management Proceedings*, *1986*(1), 118–122.

Epstein, J. L. (1988). Effective schools or effective students: Dealing with diversity. In R. Haskins & D. MacRae (Eds.), *Policies for America's public schools: Teachers, equity, and indicators* (pp. 89–126). Norwood, NJ: Ablex.

Fahy, P. J. (2006). Addressing some common problems in transcript analysis. *The International Review of Research in Open and Distributed Learning*, *1*(2), 1–6.

Farrell, C., Wohlstetter, P., & Smith, J. (2012). Charter management organizations: An emerging approach to scaling up what works. *Educational Policy*, *26*, 499–532.

Fiol, C. M., & Lyles, M. A. (1985). Organizational learning. *The Academy of Management Review*, *10*, 803–813.

Glazerman, S., & Seifullah, A. (2012). *An evaluation of the Chicago teacher advancement program (Chicago TAP) after four years. Final report*. Chicago, IL: Mathematica Policy Research.

Goldsmith, M., Morgan, H., & Ogg, A. J. (Eds.). (2004). *Leading learning organization: Harnessing the power of knowledge*. San Francisco, CA: Jossey-Bass.

Gross, B. (2011). *Inside charter schools: Unlocking doors to student success*. Seattle: Center on

Reinventing Public Education, University of Washington.

Hallinger, P., & Leithwood, K. (1994). Exploring the effects of principal leadership. *School Effectiveness and School Improvement*, *5*, 206–218.

Hamilton, L. S., Stecher, B., Marsh, J., McCombs, J. S., Robyn, A., Russell, J. L., . . . Barney, H. (2007). *Standards-based accountability under No Child Left Behind: Experiences of teachers and administrators in three states*. Santa Monica, CA: RAND.

Hamilton, L. S., Steiner, E. D., Holtzman, D., Fulbeck, E. S., Robyn, A., Poirier, J., & O'Neil, C. (2014). *Using teacher evaluation data to inform professional development in the intensive partnership sites*. Santa Monica, CA: RAND. Retrieved from http://192.5.14.43/content/dam/rand/pubs/working_papers/WR1000/WR1033/RAND_WR1033.pdf

Heck, R. (1993). School context, principal leadership, and achievement: The case of secondary schools in Singapore. *Urban Review*, *25*, 151–166.

Heneman, H. G., & Milanowski, A. T. (2003). Continuing assessment of teacher reactions to a standards-based teacher evaluation system. *Journal of Personnel Evaluation in Education*, *17*, 173–195.

Hentschke, G. C., & Wohlstetter, P. (2004). *Cracking the code of accountability* (USCUrban ed.). Los Angeles: Rossier School of Education, University of Southern California.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*, 56–64.

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal*, *48*, 794–831.

Hitt, D. H., & Tucker, P. D. (2016). Systematic review of key leader practices found to influence student achievement: A unified framework. *Review of Educational Research*, *86*, 531–569.

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel* (Research paper, MET Project). Seattle, WA: Bill & Melinda Gates Foundation.

Honig, M. I. (2003). Building policy from practice: District central office administrators' roles and capacity for implementing collaborative education policy. *Educational Administration Quarterly*, *39*, 292–338.

Honig, M. I. (2012). District central office leadership as teaching: How central office administrators support principals' development as instructional leaders. *Educational Administration Quarterly*, *48*, 733–774.

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics*, *26*, 101–136.

Jennings, J. L. (2012). The effects of accountability system design on teachers' use of test score data. *Teachers College Record*, *114*(11), 1–23.

Jiang, J. Y., Sporte, S. E., & Luppescu, S. (2015). Teacher perspectives on evaluation reform: Chicago's REACH students. *Educational Researcher*, *44*, 105–116.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains* (Research paper, MET Project). Seattle, WA: Bill & Melinda Gates Foundation.

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011). Identifying effective classroom practices using student achievement data. *Journal of Human Resources*, *46*, 587–613.

Kenny, D. (2012, October 14). Want to ruin teaching? Give ratings. *The New York Times*. Retrieved from http://www.nytimes.com/2012/10/15/opinion/want-to-ruin-teaching-give-ratings.html

Kerr, S. (1975). On the folly of rewarding A, while hoping for B. *The Academy of Management Journal*, *18*, 769–783.

Kimball, S. M. (2002). Analysis of feedback, enabling conditions and fairness perceptions of teachers in three school districts with new standards-based evaluation systems. *Journal for Personnel Evaluation in Education*, *16*, 241–268.

King, M. B., & Bouchard, K. (2011). The capacity to build organizational capacity in schools. *Journal of Educational Administration*, *49*, 653–669.

Knight, J., & van Nieuwerburgh, C. (2012). Instructional coaching: A focus on practice. *Coaching: An International Journal of Theory, Research & Practice*, *5*, 100–112.

Lake, R., Dusseault, B., Bowen, M., Demeritt, A., & Hill, P. (2010). *The national study of charter management organization (CMO) effectiveness: Report on interim findings*. Seattle, WA: Mathematica Policy Research and Center on Reinventing Public Education.

Lane, P. J., & Lubatkin, M. (1998). Relative absorptive capacity and interorganizational learning. *Strategic Management Journal*, *19*, 461–477.

Lee, J. (2010, July 19). Speeding up the race to the top. *The White House Blog*. Retrieved from http://www.whitehouse.gov/blog/2010/01/19/speeding-race-top

Leithwood, K. (2006). *Teacher working conditions that matter: Evidence for change*. Toronto, Canada: Elementary Teachers' Federation of Ontario. Retrieved from http://www.etfo.ca/

Resources/ForTeachers/Documents/TeacherWorkingConditionsThatMatter-EvidenceforChange.pdf

Leithwood, K., Leonard, L., & Sharratt, L. (1998). Conditions fostering organizational learning in schools. *Educational Administration Quarterly*, *34*, 243–276. doi:10.1177/0013161X98034002005

Leithwood, K. A., Louis, K. S., Anderson, S., & Wahlstrom, K. (2004). *Review of research: How leadership influences student learning*. Minneapolis: University of Minnesota and University of Toronto.

Leonhardt, D. (2013, October 17). Study finds gains from teacher evaluation. *The New York Times*. Retrieved from http://economix.blogs.nytimes.com/2013/10/17/a-new-look-at-teacher-evaluations-and-learning/

Levitt, B., & March, J. G. (1988). Organizational learning. *Annual Review of Sociology*, *14*, 319–340.

Lloria, M. B. (2007). Differentiation in knowledge-creating organizations. *International Journal of Manpower*, *28*, 674–693.

Louis, K. S. (2007). Changing the culture of schools: Professional community, organizational learning, and trust. *Journal of School Leadership*, *16*, 477–487.

Louis, K. S., & Lee, M. (2016). Teachers' capacity for organizational learning: The effects of school culture and context. *School Effectiveness and School Improvement*, *27*, 534–556. doi:10.1080/09243453.2016.1189437

Louisiana Department of Education. (2013). *Louisiana teacher performance evaluation rubric*. Baton Rouge: Author.

Louisiana Department of Education. (2015). *Ten years after Hurricane Katrina: The New Orleans educational landscape today*. Retrieved from https://www.louisianabelieves.com/resources/about-us/10-years-after-hurricane-katrina

Lubienski, C. (2003). Innovation in education markets: Theory and evidence on the impact of competition and choice in charter schools. *American Educational Research Journal*, *40*, 395–443.

March, J. G. (1994). *Primer on decision making: How decisions happen*. New York, NY: Free Press.

Marsh, J., Bertrand, M., & Huguet, A. (2015). Using data to alter instructional practice: The mediating role of coaches and professional learning communities. *Teachers College Record*, *117*(4), 1–40.

Marsh, J., McCombs, J. S., Lockwood, J. R., Martorell, F., Gershwin, D., Naftel, S., Le, V., Shea, M., Barney, H., & Crego, A. (2008). *Supporting literacy across the sunshine state: A study of Florida middle school reading coaches*. Santa Monica, CA: RAND, MG-762-EDU.

Massell, D. (1998). *State strategies for building local capacity: Addressing the needs of standards-based reform* (CPRE Policy Briefs). Philadelphia, PA: Consortium for Policy Research in Education.

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

McLaughlin, M. W., & Talbert, J. E. (1993). *Contexts that matter for teaching and learning*. Stanford, CA: Center for Research on the Context of Secondary School Teaching, Stanford University.

McLaughlin, M. W., & Talbert, J. E. (2001). *Professional communities and the work of high school teaching*. Chicago, IL: The University of Chicago Press.

Milanowski, A. T., & Heneman, H. G. (2001). Assessment of teacher reactions to a standards-based teacher evaluation system: A pilot study. *Journal of Personnel Evaluation in Education*, *15*, 193–212.

Miles, M. B., Huberman, A. M., & Saldaña, J. (2013). *Qualitative data analysis: A methods sourcebook*. Thousand Oaks, CA: SAGE.

Mintrop, R. (2012). Bridging accountability obligations, professional values, and (perceived) student needs with integrity. *Journal of Educational Administration*, *50*, 695–726.

Murphy, J., Hallinger, P., & Heck, R. H. (2013). Leading via teacher evaluation: The case of the missing clothes? *Educational Researcher*, *42*, 349–354.

Newmann, F. M., King, M. B., & Youngs, P. (2000). Professional development that addresses school capacity: Lessons from urban elementary schools. *American Journal of Education*, *108*, 259–299.

O'Day, J. A. (2002). Complexity, accountability, and school improvement. *Harvard Educational Review*, *72*, 293–329.

O'Day, J. A., Goertz, M. E., & Floden, R. E. (1995). *Building capacity for education reform* (Vol. 18). New Brunswick, NJ: Consortium for Policy Research in Education.

Pintrich, P. R. (2003). A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of Educational Psychology*, *95*, 667–686.

Preston, C., Goldring, E., Berends, M., & Cannata, M. (2012). School innovation in district context: Comparing traditional public schools and charter schools. *Economics of Education Review*, *31*, 318–330.

Purkey, S. C., & Smith, M. S. (1983). Effective schools: A review. *The Elementary School Journal*, *83*, 427–452.

Robinson, V. M., Lloyd, C. A., & Rowe, K. J. (2008). The impact of leadership on student outcomes: An analysis of the differential effects of leadership types. *Educational Administration Quarterly*, *44*, 635–674.

Rondinelli, D. (1981). Government decentralization in comparative perspective: Theory and practice in developing countries. *International Review of Administrative Sciences*, *47*, 133–145.

Rondinelli, D. (1989). Decentralizing public services in developing countries: Issues and opportunities. *Journal of Social, Political, and Economic Studies*, *14*, 77–98.

Sartain, L., Stoelinga, S. R., & Brown, E. (2009). *Evaluation of the excellence in teaching pilot: Year 1 report to the Joyce Foundation*. Chicago, IL: The Consortium on Chicago School Research at the University of Chicago.

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* (Research report). Chicago, IL: University of Chicago Urban Education Institute. Retrieved from http://files.eric.ed.gov/fulltext/ED527619.pdf

Schön, D. A. (1987). *Educating the reflective practitioner: Toward a new design for teaching and learning in the professions*. San Francisco, CA: Jossey-Bass.

Schraw, G. (1998). Promoting general metacognitive awareness. *Instructional Science*, *26*, 113–125.

Scott, R. W. (1998). *Organizations: Rational, natural, and open systems*. Upper Saddle River, NJ: Prentice Hall.

Shallcross, D. J. (1975). Creativity: Everybody's business. *Personnel and Guidance Journal*, *51*, 623–626.

Shipton, H., Dawson, J., West, M., & Patterson, M. (2002). Learning in manufacturing organizations: What factors predict effectiveness? *Human Resource Development International*, *5*, 55–72.

Smylie, M. A. (2009). *Continuous school improvement*. Thousand Oaks, CA: Corwin Press.

Spillane, J. P., & Louis, K. S. (2002). School improvement processes and practices: Professional learning for building instructional capacity. *Yearbook of the National Society for the Study of Education*, *101*, 83–104.

Spillane, J. P., & Miele, D. B. (2007). Evidence in practice: A framing of the terrain. *Yearbook of the National Society for the Study of Education*, *106*, 46–73.

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

Spillane, J. P., & Thompson, C. L. (1997). Reconstructing conceptions of local capacity: The local education agency's capacity for ambitious instructional reform. *Education Evaluation and Policy Analysis*, *19*, 185–203.

Steinberg, M. P., & Garrett, R. (2016). Classroom composition and measured teacher performance: What do teacher observation scores really measure? *Educational Evaluation and Policy Analysis*, *38*, 293–317.

Stoll, L. (1999). Realising our potential: Understanding and developing capacity for lasting improvement. *School Affectiveness and School Improvement*, *10*, 503–532.

Stoll, L. (2009). Capacity building for school improvement or creating capacity for learning? A changing landscape. *Journal of Educational Change*, *10*, 115–127.

Strunk, K. O., Weinstein, T. L., & Makkonnen, R. (2014). Sorting out the signal: Do multiple measures of teachers' effectiveness provide consistent information to teachers and principals. *Education Policy Analysis Archives*, *22*, 1–41.

Suh, T. (2002). Encouraged, motivated and learning oriented for working creatively and successfully: A case of Korean workers in marketing communications. *Journal of Marketing Communications*, *8*, 135–147.

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

Tennessee Department of Education. (2012). *Teacher evaluation in Tennessee: A report on year 1 implementation*. Nashville: Author.

Urdan, T., & Schoenfelder, E. (2006). Classroom effects on student motivation: Goal structures, social relationships, and competence beliefs. *Journal of School Psychology*, *44*, 331–349.

U.S. Department of Education Office of Inspector General. (2011, December). *Department's implementation of the Teacher Incentive Fund grant program: Final audit report*. Retrieved from http://www2.ed.gov/about/offices/list/oig/auditreports/fy2012/a19i0007.pdf

Waters, T., Marzano, D. R. J., & McNulty, B. (2003). *Balanced leadership*. Aurora, CO: McREL.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect* (2nd ed.). The New Teacher Project. Retrieved from http://tntp.org/assets/documents/TheWidgetEffect_2nd_ed.pdf

Woodman, R. W., Sawyer, J. E., & Griffin, R. W. (1993). Toward a theory of organizational creativity. *The Academy of Management Review*, *18*, 293–321.

Yin, R. K. (2013). *Case study research: Design and methods*. Thousand Oaks, CA: SAGE.

## Authors

JULIE A. MARSH, PhD, is an associate professor of education policy at the University of Southern

California's Rossier School of Education, who specializes in research on K–12 policy. Her research blends perspectives in education, sociology, and political science. Her research focuses on the implementation and effects of accountability and instructional reform policies, including the roles of central office administrators, intermediary organizations, and community members in educational reform and the use of data to guide decision making.

SUSAN BUSH-MECENAS is a PhD candidate and Provost's fellow at the University of Southern California's Rossier School of Education. Her research interests include organizational learning, district reform, district and school capacity building, and accountability policy.

KATHARINE O. STRUNK, PhD, is an associate professor of education and policy at the University of Southern California. Her research centers on K–12 education policy with a focus on education labor markets, accountability policies, and governance. In particular, she studies teachers' unions and the collective bargaining agreements they negotiate with district administrators and the implementation and impacts of various teacher retention, evaluation, and compensation policies, as well as broader accountability policies.

JANE ARNOLD LINCOVE is an associate professor at the School of Public Policy, University of Maryland and a research fellow at the Education Research Alliance for New Orleans at Tulane University. Her research focuses on the implementation and effects of market-based policy in public education, economics of education, and equity.

ALICE HUGUET is a postdoctoral fellow at Northwestern University's School of Education and Social Policy. Her research explores school- and district-level responses to state and federal policy in varied contexts. She is interested in organizational arrangements and how data and research are used in decision-making processes at these levels.

Manuscript received July 10, 2016
First revision received November 4, 2016
Second revision received January 31, 2017
Accepted February 1, 2017