Can Principals Promote Teacher Development as Evaluators?
A Case Study of Principals' Views and Experiences

Matthew A. Kraft
Brown University

Allison Gilmour
Vanderbilt University

Abstract

Purpose: New teacher evaluation systems have expanded the role of principals as instructional leaders, but little is known about principals' ability to promote teacher development through the evaluation process. We conducted a case study of principals' perspectives on evaluation and their experiences implementing observation and feedback cycles to better understand whether principals feel as though they are able to promote teacher development as evaluators.

Research Methods: We conducted interviews with a stratified random sample of 24 principals in an urban district that recently implemented major reforms to its teacher evaluation system. We analyzed these interviews by drafting thematic summaries, coding interview transcripts, creating data-analytic matrices, and writing analytic memos.

Findings: We found that the evaluation reforms provided a common framework and language that helped facilitate principals' feedback conversations with teachers. However, we also found that tasking principals with primary responsibility for conducting evaluations resulted in a variety of unintended consequences which undercut the quality of evaluation feedback they provided. We analyze five broad solutions to these challenges: strategically targeting evaluations, reducing operational responsibilities, providing principal training, hiring instructional coaches, and developing peer evaluation systems.

Implications: The quality of feedback teachers receive through the evaluation process depends critically on the time and training evaluators have to provide individualized and actionable feedback. Districts that task principals with primary responsibility for conducting observation and feedback cycles must attend to the many implementation challenges associated with this approach in order for next-generation evaluation systems to successfully promote teacher development.

Keywords: teacher evaluation, principals, teacher development, class observation, feedback

District- and state-level efforts to remake teacher evaluation systems are among the most widely adopted reforms that U.S. public schools have experienced in decades (McGuinn, 2012). These reforms were motivated in large part by research documenting that teachers have large effects on student learning (Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004), and that existing evaluation systems were perfunctory and narrowly focused on compliance (Tucker, 1997; Weisberg, Sexton, Mulhern, & Keeling, 2009). The Obama administration has sought to strengthen teacher quality by making teacher evaluation reforms the centerpiece of its signature education initiative, Race To The Top (RTTT), as well as state-waivers to No Child Left Behind (NCLB). Today, 46 states have enacted new legislation aimed at strengthening and expanding teacher evaluation systems in public schools (Steinberg & Donaldson, in press).

Research on these next-generation evaluation systems has focused overwhelmingly on policy goals, program designs, and performance measures (e.g. Kane, McCaffrey, Miller, & Staiger, 2013). However, we still know very little about how these policies are interpreted and enacted by school leaders. History clearly shows that the success of federal, state, and local policy initiatives depends on the will and capacity of local actors to implement reforms (Honig, 2006). This is particularly true in the decentralized U.S. education system where local practice is often decoupled from central policy (Spillane & Kenney, 2012).

In this case study, we examine the perspectives and experiences of the local actors who are primarily responsible for implementing evaluations – school principals. School principals have supervised and evaluated teachers for well over a century (Cubberley, 1916). In keeping with this tradition, many states and districts require principals to conduct observation and feedback cycles as part of new evaluation systems (Center on Great Teachers and Leaders, 2014;

Herlihy et al., 2014). In a number of states, including the one in which our study takes place, principals are given full responsibility for determining teachers' overall summative evaluation ratings (Donaldson & Papay, 2014; Steinberg & Donaldson, in press).

Relying on principals as the primary evaluators raises important questions about their willingness, capacity, and ability to implement observation and feedback cycles and support teacher development through the evaluation process. However, we know very little about principals' perspective on the evaluation process. Some scholars (Hanushek, 2009) and journalists (Thomas, Wingert, Conant, & Register, 2010) see evaluation as a mechanism for increasing teacher effort through accountability and monitoring, and for dismissing ineffective teachers. Others view evaluation as a process that can support the professional growth of all teachers by promoting self-reflection, establishing a common framework for analyzing instruction, and providing individualized feedback (Almy, 2011; Curtis & Wiener, 2012). On paper, policymakers privilege this latter view; nearly every state identified professional learning as the primary purpose of evaluation reforms in their NCLB waiver applications (Center on Great Teachers and Leaders, 2014). In practice, districts often hope to promote development while also using evaluations for high-stakes accountability (Steinberg & Donaldson, in press).

Evaluation system reforms have also greatly expanded the demands on principals' time and the role of principals as instructional leaders. For decades, principals typically completed one-time observation check-lists and then provided copies to teachers. New systems require multiple observations using extensive rubrics, detailed written feedback, and post-observations meetings to provide feedback (Danielson, 2007; Stronge, 2005). The degree to which principals are prepared to assume this expanded role and the ways in which they navigate these responsibilities have important implications for teacher development (Lavigne & Good, 2015).

We explored these issues by interviewing principals from a large urban school district in the northeastern United States that had implemented reforms to its teacher evaluation system. We conducted interviews with 24 district principals recruited to participate using a stratified random sampling design. We interviewed principals in the summer after the first year of district-wide implementation of the redesigned teacher evaluation system. The first year the district did not use any measures of teacher effectiveness based on student achievement tests. This allowed us to understand principals' experiences with the observation portion of teacher evaluations without confounding these experiences with the controversy surrounding value-added measures.

Our case study focuses on principals' perspectives and experiences with classroom observation and feedback because this process is a primary mechanism through which evaluation is intended to promote teacher development. Principals' abilities to rate teachers accurately, to facilitate teachers' own self-reflection, to make specific actionable recommendations, and to communicate this feedback effectively are central to any evaluation process intended to improve instruction. This paper makes several contributions to the literature. First, the paper is among the first to look inside the black box of how next-generation evaluations systems are perceived and implemented by principals. Second, we describe how, in the district we studied, four key implementation challenges resulted in unintended consequences that undercut principals' ability to support teachers' professional growth. Finally, the paper discusses five different proposals to improve the quality of feedback teachers receive through observation and feedback cycles.

Teacher Evaluation Reforms: Theory and Implementation

Teacher Evaluation Feedback for Professional Improvement

The purpose of teacher evaluation is, in theory, twofold: to serve as a professional development process and as a quality assurance mechanism (Danielson & McGreal, 2000).

Historically, teacher evaluation systems have rarely served either aim. Evaluation systems did not differentiate among teachers, were rarely used to inform personnel decisions, and failed to provide meaningful feedback to teachers (Toch & Rothman, 2008; Tucker, 1997; Weisberg et al., 2009). These findings combined with federal initiatives have spurred widespread reforms to the design of teacher evaluation systems at the state and local levels (Donaldson & Papay, 2014). New systems now commonly incorporate multiple measures of teacher performance and rate teachers across multiple performance categories (Steinberg & Donaldson, in press).

Efforts to leverage the evaluation process as a professional development tool are centered on the classroom observation process. Rating of teachers' instructional practice on classroom observation rubrics are now a universal feature of new evaluation systems (Steinberg & Donaldson, in press). The theory of action for how observation and feedback cycles can promote professional growth includes several mechanisms (Curtis & Weiner, 2012; Papay, 2012). First, observation rubrics provide teaches with a common framework for planning, enacting, and discussing classroom instruction. Second, the observation and feedback process can develop teachers' habits and abilities to reflect on their own practices and assess their own strengths and weaknesses. Third, evaluators can provide teachers with specific and actionable feedback on how they might improve their instructional practice or serve as a sounding board as teachers drive their own improvement process. Finally, the observation and feedback process provides a formal structure that pushes teachers to set goals and tracks their progress towards meeting these goals.

These cycles of observation, reflection, dialogue and feedback, and goal setting can provide teachers with new ideas as well as frequent and relevant feedback to support their professional growth. A key assumption of this theory of action is that teachers are both willing and able to improve their practice by actively engaging in the evaluation process. No amount of

5

feedback will result in professional growth if a teacher is unwilling or unable to co-construct and enact changes. The literature largely supports this assumption, documenting both teachers' willingness (Kennedy, 2005) and ability (Authors, 2014) to improve their practice over time.

Many scholars and practitioners view these rubric-based assessments, and subsequent conversations between evaluators and teachers, as providing new opportunities to foster professional development at scale (Almy, 2011, Donaldson & Peske, 2010). However, much of the initial focus of implementing new evaluation systems has focused on the design features of the assessment process: selecting performance measures, developing information management systems, standardizing observation procedures, and determining weights and score threshold to map multiple performance ratings onto a single performance evaluation category (Kane, Kerr, & Pianta, 2014). Investments in personnel training and protocols for supporting professional development through the evaluation process have been far more limited (Lavigne & Good, 2015).

Implementing observation and feedback cycles as part of high-stakes evaluation systems also presents a range of implementation challenges. The polemic and personal nature of teacher evaluation combined with the resources it requires suggests principals will confront considerable challenges and difficult tradeoffs (Halverson, Kelley, & Kimball, 2004). Using the evaluation process as a means to promote professional learning requires principals to confront perceptions among teachers that evaluation is primarily intended to dismiss low-performing teachers (Thomas, Wingert, Conant, & Register, 2010). Principals must navigate potentially conflicting assessments of teachers' effectiveness due to the relatively low correlations between scores on observation rubrics and teacher value-added measures (Hill, Kapitula, & Umland, 2011; Kane & Staiger, 2012). Inaccurate evaluations due to insufficient training, lack of time, evaluator bias, and imprecise measures can impose substantial costs by causing poor staffing decisions,

misdirecting teachers' efforts for improvement, and undercutting relational trust among school staff. If principals view new evaluation reform initiatives as under-resourced or unrealistic, they may respond by "satisficing" – focusing on compliance rather than high-quality implementation (Halverson & Clifford, 2006). How districts address these implementation challenges plays an equally important role in determining the success of evaluation reforms as the design of the systems themselves.

Principals as Instructional Leaders and Evaluators

The role and responsibilities of school principals have evolved continually over the last century in response to shifting policy landscapes and public expectations (Spillane & Kenney, 2012). Principals are at once building managers, employers, professional figureheads, supervisors, inspirational leaders, and providers of professional development. They shape the experiences of teachers and students through these interrelated roles (Leithwood & Louis, 2011; Waters, Marzano, & McNulty, 2003). The quality of principal leadership as measured by teacher surveys is a strong predictor of teacher turnover and student achievement across schools (Authors, 2012; Authors, 2015, Boyd et al., 2011; Ladd, 2011). Theoretical models and empirical evidence suggest that principal effects operate through both direct and indirect pathways (Witziers, Bosker, & Kruger, 2003). Several studies have found positive associations between principal characteristics and leadership styles and student achievement that are mediated by their influence on the school climate, instructional practices, and the quality of professional development (Hallinger, Bickman, & Davis, 1996; Sebastian & Allensworth, 2012; Supovitz, Sirinides, & May, 2010).

Principals' roles have expanded to encompass a direct role in shaping student learning via instructional leadership (Robinson, Lloyd, & Rowe, 2008; Supovitz et al., 2010). Instructional

leadership includes staff development, curriculum development and coherence, student assessment and analysis, and evaluation and individualized feedback (Hoy & Hoy, 2012; Newmann, Smith, Allensworth, & Bryk, 2001). Principals can facilitate peer learning opportunities for teachers by developing teacher teams with clear purposes, building in common planning time, and providing opportunities for peer observations and feedback (Louis, Dretzke, & Wahlstrom, 2010). They play a key role in developing a school-wide culture of high-expectations for students which is directly linked to student achievement (Authors, 2015)

Studies of principals' time use prior to new evaluation reforms suggest they spent only a small fraction of their time on instructional leadership activities. Horng, Klasik, and Loeb (2010) found that principals spent less than 6% of their time observing, coaching, and evaluating teachers and only 7% developing and delivering instructional programming. May and Supovitz's (2011) analysis revealed that principals spent an average of 8% of their time on instructional leadership activities, but that this average masked considerable heterogeneity. Grissom, Loeb, and Master (2013) found that principals spent less than 13% of their time on instructional activities.

New teacher evaluation system reforms have greatly expanded principals' instructional leadership responsibilities by requiring principals to work one-on-one with teachers to evaluate and improve their classroom practices. While it is clear that new evaluation systems require that principals take on expanded roles as instructional leaders, we know less about how they are managing these responsibilities or the results of their efforts. Halverson, Kelley, and Kimball's (2004) analysis of the school-level implementation of a new observation system found that the system consumed as much as 25% of principals' time and resulted in satisficing behaviors such as brief observations and positive generic feedback. The absence of formative or critical

feedback in written evaluations led them to conclude that "evaluators lacked the skills to provide valuable feedback, particularly with accomplished teachers" (p. 178). Similarly, Sartain, Stoelinga, and Brown (2011) studied the pilot of a new evaluation system in Chicago Public Schools and found that principals spoke about 75% of the time during conferences and only 10% of their questions were higher-order questions that pushed teachers to reflect. Sartain and her colleagues concluded that "principals need more support in engaging in deep coaching conversations" (p. 21). Other studies further suggest that principals face substantial capacity constraints (Donaldson, 2012; Donaldson, 2013)

Despite these challenges, there is some evidence that evaluation systems with principals as evaluators may help improve teacher effectiveness. Steinberg and Sartain (in press) exploit CPS's randomized rollout of a new pilot evaluation system to estimate the causal effect of evaluation on student achievement. The authors found that the new evaluation system produced significant improvements in reading achievement and positive, but imprecisely estimated, effects in mathematics. However, the authors found no effect in either subject among the cohort of schools who adopted the system in the second year, possibly due to the reduction in training and support for principals in the second year. Taylor and Tyler (2012) analyzed an evaluation program in Cincinnati Public Schools in which teachers were observed by peer evaluators three times and by principals once. Peer evaluators were high-performing teachers from other schools in the district who completed training on the new evaluation system. The authors found that frequent observation and feedback cycles with peer evaluators as well as principals raised student achievement in mathematics, but found no effect on reading achievement.

Taken together, these studies suggest that there is potential for high-quality observation and feedback cycles to promote teacher development, but that it remains unclear whether

principals have the time, training, and support necessary to implement these cycles effectively. We build on this body of literature by exploring the implications of relying on principals to conduct observation and feedback cycles as part of next-generation evaluation systems with a focus on the following questions: (1) What are principals' views on the purpose of teacher evaluation? (2) How do principals balance their expanded roles as instructional leaders with their other responsibilities? (3) What are principals' experiences implementing observation and feedback cycles? (4) What are principals' perspectives on how to improve the quality of feedback teachers' receive through the evaluation process?

The District Evaluation System in Context

The former evaluation system used by the district we studied was, in many ways, typical of those characterized in The Widget Effect report (Weisberg et al., 2009). The system stipulated that administrators should rate new teachers annually and permanent teachers biannually using a rubric with a binary rating scale. Teachers received an overall rating as well as ratings on eight different dimensions of professional practice. Principals were required to write an individualized improvement plan for any teachers receiving an overall rating of unsatisfactory. If the teacher failed to improve, the principal was required to write a second improvement plan and could initiate the dismissal process. Moving towards dismissal meant following a strict timeline of interim observations that could take up to two years to complete.

Studies of the former evaluation system in the district suggest that it was more a perfunctory process than a useful tool for promoting teacher development or dismissing ineffective teachers.[1] An analysis of the district evaluation process by an independent non-profit organization found that evaluations were superficial and infrequent; many teachers went unevaluated and schools often failed to submit the required evaluations to the district. A report

by the state teachers' union argued the extensive evaluation checklist was too complicated with almost 20 behavioral statements and 72 indicators which did not lend themselves easily to observation or measurement. In light of these weaknesses, the district implemented a new evaluation system in 2011 that was built on the state's new evaluation regulations and adapted for the district's context in partnership with the local teacher's union.

The current evaluation system in the district shares many features that are common across states and districts which have implemented major reforms to their evaluation practices. In the year leading up to the full-scale rollout of this current system, principals and other evaluators received in-depth training intended to familiarize them with the features of the new system and calibrate their classroom observation ratings of teachers' performance. The district was explicit about its intent to shift the purpose and perception of evaluation from compliance to teacher development, emphasizing it was "designed first and foremost to promote leaders' and teachers' growth and development." The evaluation process is centered on a continuous cycle of assessment using an original rubric developed by the state and adapted by the district that captures observable standards related to teaching effectiveness. This rubric is comprised of four broad domains capturing Curriculum Design and Assessments, Instructional Practice, Family Engagement, and Professionalism. Each of these domains consists of between three to six indicators with a total of 34 distinct elements that teachers are rated using a four-point scale.

Principals and select members of their administrative teams (e.g. Assistant Principals, Directors of Instruction) are responsible for providing teachers with a mid-year formative assessment and an end-of-year summative assessment. Assessments include an overall rating, ratings on each rubric domain, and evaluations of their progress towards achieving professional practice (PP) and student learning (SL) goals. Teachers are active participants in the evaluation

process; they initiate each cycle by self-assessing their own work and designing action plans to achieve PP and SL goals. Evaluators conduct 1 to 4 formal unannounced observations of each teacher throughout the year, depending on a teacher's prior evaluation rating, and provide formal written feedback after each observation. In addition, evaluators are encouraged to conduct frequent informal observations lasting 15-20 min and hold face-to-face post-observation conversations with teachers. Evaluators use evidence from classroom observations and artifacts submitted by teachers documenting their progress towards professional practice and student learning goals to inform their ratings. Teachers rated in the top two categories continue this cycle of self-directed growth while those in the lower rating categories are placed on more structured evaluation plans, which, after several repeated low evaluations, can result in dismissal.

Many of the core features of the district's current system are common across next-generation teacher evaluation systems adopted by states and districts. In their comprehensive review of recent teacher evaluation reforms, Steinberg and Donaldson (in press) found that all 46 states which have implemented reforms have designated classroom observation ratings as the central evaluation measure. Over half of all states also include student learning objectives where teachers develop goals for what students should achieve and assess students' progress towards these goals (Lacireno-Paquet, Morgan, & Mello, 2014).

Although data on how districts and states implement these systems is less readily available, existing evidence suggests that districts commonly task principals with the responsibility of evaluating teachers. Many urban districts including Chicago, Los Angeles, Miami-Dade, New York City, and Washington, D.C. require principals to conduct classroom observations.[2] At the state level, many systems require principals, assistant principals, or other administrators to conduct evaluations (Center on Great Teachers and Leaders, 2013). Based on

their analysis of interviews with state education officials and evaluation system documents from 17 states, Herlihy and her colleagues (2014) concluded that most new evaluation systems appeared to default to the past approach where principals served as the sole evaluator (p. 14). Among state applications for RTTT funds, we find that 22 states identified principals, administrators, or school leaders as responsible for conducting observations, while nine referenced "trained evaluators" and the remaining eight did not specify who would conduct observations.

One key difference between the district's approach and most other systems is that it places the responsibility of arriving at an overall rating squarely on the shoulders of principals. Steinberg and Donaldson (in press) found that only 14 of 46 states took a similar approach of requiring evaluators to consider all evidence and make final summative judgements. Instead, most states specify a formula for arriving at an overall score based on the weighted sum of multiple evaluation measures. This feature of the evaluation system further amplifies the consequential weight of the evaluation responsibilities principals' carried in the district.

<div align="center">**Research Methods**</div>

Sample

The district we studied is an urban district in the northeast that serves a racially and linguistically diverse student population. Hispanic and African American students make up approximately 75% of the district student body, while the remaining 25% of students are predominantly Caucasian and Asian American. Over 70% of students in the district are eligible for free or reduced price lunch and nearly half speak a language other than English as their first language. We defined our target population of inference as all principals in the district that oversaw schools serving students in main-stream classes across grades K-12. This included

traditional district schools, exam schools which admit students based on standardized test scores, and semi-autonomous district schools that have autonomy over budget, staffing, governance, curriculum/assessment, and the school calendar. We purposely excluded early childhood centers, vocational and technical schools, and alternative schools for students with disabilities.

Early in the summer of 2013, we recruited a subset of 46 randomly selected principals to participate in the study in order to capture views that were broadly representative of principals across the district as a whole. In order to reduce chance sampling idiosyncrasies that might skew our results, we identified potential participants using a stratified random sampling framework. We chose two school characteristics, school size and level, on which to stratify our sample. Specifically, we categorized all principals into six different strata: three school types (elementary, middle, and high) and two school sizes (390 students or more, less than 390 students). We then contacted up to nine randomly selected principals within each strata by phone and email to invite them to participate confidentially in our study.

Our sampling procedure resulted in a diverse collection of interview participants with demographic characteristics and school assignments that were broadly representative of the district as a whole. Twenty-four out of the 46 principals we contacted agreed to be interviewed, a participation rate of 52%. Ten of the participating principals were African American, eight were Caucasian, two were Asian American, two were Hispanic and two were of mixed race. Figure 1 Panel A illustrates the range of prior teaching experience among the sample. All principals except one had prior experience in the classroom with an average of just under 10 years across the sample. Administrative experience varied across the sample with an average of just over 10 years of total experience as administrators. However, Figure 1 Panel B illustrates how most principals were relatively new to the schools where they currently worked.

We conducted a series of t-tests to confirm that our stratified random sample of participating principals is representative of principals across the district. In Table 1, we provided the demographic characteristics and school characteristics for all principals in the district we interviewed and those we did not. We find no statistically significant differences across any measures, strong evidence that our sample is broadly representative of the district as a whole.

The principals we spoke with worked across the full range of school types, levels, and sizes. Our sample included principals of 15 traditional district schools, six semi-autonomous schools, two exam schools, and one in-district charter school. These schools varied by levels and size: five small and six large elementary schools, three small and three large middle schools, and two small and five large high schools. School size in the district is closely related to the number of administrators who were authorized to conduct teacher evaluations at a school. At nine of the smallest schools in our sample, principals were the only evaluators. Principals at nine other medium-sized schools had one or two other administrators who also conducted evaluations, while the five largest middle and high schools had 3 to 9 additional evaluators.

The student populations in the schools where participating principals worked ranged widely and closely mirrored the distribution of student body characteristics across all schools in the district. For example, the percentage of students scoring proficient on mathematics state exams in 4th through 8th and 10th grade ranged from 16% to 96%. Four schools has less than 25% of students score proficient in math, 13 schools had between 25% and 50% score proficient, four schools had between 50% and 75% proficient, and 3 schools had over 75% score proficient. The variability in English language arts proficiency rates closely mirrors that of math.

Data Collection and Analysis

We conducted interviews with principals lasting 45 to 60 minutes in July and August of 2013, the summer after the first year the new evaluation system was implemented district-wide. These interviews gave principals the opportunity to share their perspectives about teacher evaluation as well as their experiences implementing the districts' former and current evaluation systems. The authors and a research assistant conducted each interview individually in person, or by phone, based on principals' availability and preferences. We used a semi-structured protocol (see Appendix A) to ensure that each interview touched upon a common set of topics and reduced interviewer effects and bias (Patton, 2001). We audio-recorded each conversation and transcribed the interviews to facilitate data analysis. Our research team then composed structured, thematic summaries (Maxwell, 2005) of each interview and used these summaries to develop a set of codes that captured the common themes and topics raised by principals.

We coded interview transcripts for central concepts (Strauss & Corbin, 1998) using a hybrid approach to developing codes (Miles & Huberman, 1994). We generated codes informed by our research questions, the theory of action behind classroom observation and feedback cycles, and our review of the instructional leadership literature discussed above, as well as common topics that were reflected in our thematic summaries. Each author then conducted a trial coding process with two transcripts, reviewed the other's initial coding, and debriefed about coding discrepancies and common themes that were not included in our initial set of codes. This peer-review process served to calibrate our coding approach and revealed how some of our original codes were too narrowly focused. We then iteratively refined and revised codes as new ideas emerged from the data, frequently returning to transcripts for multiple rounds of coding (see Appendix B for our original and final codes). We analyzed our interview data by organizing codes around broad themes and reviewing interview passages associated with the codes. We

wrote analytic memos that outlined the range of perspectives and experiences that principals

shared, and reviewed the characteristics of principals and their schools to situate quotes within

context. Once the evidence on each theme was organized into an extended analytic memo, we

returned to the interview transcripts to search for disconfirming evidence and counterexamples.

Findings

Evaluation Reforms Provided an Improved System for Promoting Teacher Development

While principals were candid about the limitations of the current evaluation system as it

was being first implemented in the district, all principals cited meaningful ways in which the

current system was an improvement over the former system. Three key reforms enhanced the

likelihood that principals could use the evaluation process to support teacher develop. Some of

these reforms such as the new evaluation rubric supported principals in specific and direct ways.

Others such as expanding teachers' roles in the evaluation process and shifting the evaluation

culture served to support principals in more indirect ways by facilitating the feedback process.

Evaluation rubric provided common language and specific assessments. Nearly 70%

of our sample reported that the new evaluation rubric was an important, and positive,

improvement to the new evaluation system. These principals felt that ratings based on observable

teacher practices catalogued on the rubric elements helped teachers understand why they

received certain feedback, making the evaluation process seem less subjective. The language

used on the rubric was easy for principals and teachers to understand. As one principal said:

> The language of the rubric clearly spells out what is exemplary; what is proficient; what
> is needs improvement and what is unsatisfactory. It's pretty clear what you're seeing and
> kind of which box on the rubric something's gonna fit into.

The benefit of this increased clarity was echoed by others including a principal at a large middle

school, "I definitely feel like the rubric has focused us a lot more on kind of common

understanding and common language of what we ought to be seeing. Then as I'm providing feedback, it's able to be linked to that language." For most principals, the common framework about what professional practices the district prioritized and what exemplary practice looked like provided helpful structure for their feedback conversations.

The transition from binary ratings to a rubric with four performance levels also helped principals to provide more specific feedback as part of the evaluation process. As one young administrator of a large high school explained, "the new system, because it has a bigger range, allows you to more narrowly define where they're unsatisfactory, in a more productive way." An experience middle-aged administrator also found that the new rubric pushed him to improve his feedback. He described how the expanded rubric "is extremely helpful in forcing me to, and encouraging me to, be precise with people about what they need to work on." The evaluation system structures, such as the rubric, directly shaped how principals executed feedback cycles. The shared language between administrators and teachers and specific feedback facilitated by the rubric were important features of the theory of action behind evaluation and feedback cycles.

Teachers' active role in evaluation. When asked about their views on the new evaluation system, all but eight principals cited the increased involvement of teachers in the evaluation process as an important change. As part of the new system, teachers are required to identify and work towards professional practice and student learning goals and submit artifacts to evidence their performance. According to a middle school principal with 12 years of experience, the new process "gives teachers much more control." Expectations for conducting post-observation meetings with teachers also created opportunities for teachers to engage in a productive dialogue about their performance. One principal described how the new system "offers an opportunity to really have that back and forth with people." Two principals noted that

some teachers were also supporting their peers to improve on their formative ratings. One principal described a team of teachers who worked together to help improve practice:

> When you have proficient and exemplary teachers working together—even with the needs improvement teacher or someone who had the needs improvement category. That's where we saw some really great growth in teachers working with each other.

At this exam school, teachers had begun to take ownership over supporting their peers to meet their professional practice goals.

The new evaluation system created a formalized process that promoted teacher reflection and goal setting, which are central to the theory of action for promoting professional development through evaluation. Teachers were recognized for their expertise and actively engaged in the evaluation process, which may have led them to take more ownership of the evaluation process and promote professional growth.

Shifting the culture around teacher evaluation. Fourteen principals felt that transitioning from a system of infrequent evaluations with a focus on low-performing teachers to a system where all teachers were evaluated regularly on a detailed rubric had begun to shift the "gotcha" culture around evaluation. Principals perceived this change as beginning to increase teachers' willingness to engage with them in the observation and feedback process. One principal said, "I think there's definitely less of a feel around, this is going to be used as a tool to terminate teachers." As another principal put it, "The new evaluation system does not have an 'out to get you' impression." However, five principals characterized the current evaluation process as "still very formal" and teachers as being "a little bit edgy" and "still very paranoid" even though these principals all described their evaluation efforts as focused on professional growth. In the view of an elementary school principal, her staff felt the current system was still a "gotcha" system.

Principals described positive interactions with some teachers, but for others, "once you got to the evaluation part they froze because they had had such a bad [prior] experience."

The reforms to the teacher evaluation system in the district provided a strong framework for assessing and discussing teachers' professional practice. As intended by the district, teachers were becoming more involved and the culture around evaluation was beginning to focus on professional growth. These changes facilitated principals' efforts to promote growth among their staff. However, we heard time and again that placing the full responsibility of observing and coaching teachers on principals and their administrative teams resulted in a variety of unintended consequences that undercut the potential to promote growth through the evaluation process.

Implementation Challenges and the Unintended Consequences of Relying on Principals as Evaluators

Principals experienced a variety of challenges in their efforts to implement the new evaluation system and promote teacher development. Some of these were technical challenges such as coordinating observations times and navigating the new on-line evaluation system. Most principals were quick to recognize that these were transitional costs that would become less of a burden once they had developed new routines and become familiar with the new system. However, relying on principals to evaluate teachers as a central part of the new system resulted in a range of implementation challenges. These challenges led to unintended consequences which limited the effectiveness of the feedback teachers' received in several important ways.

Challenge #1: Principals' views on the purpose of evaluation differ. As the primary observers, principals were the face of the teacher evaluation system. Principals' own perspectives on evaluation directly shaped how they chose to implement the evaluation system, and ultimately, how teachers experienced the evaluation process. We found a range of perspectives

among principals about the primary purposes and value of teacher evaluation systems. We also found that principals' views on what the evaluation system should be used for did not always align with how the district articulated the purpose of the system or how principals felt teachers perceived the system. These differing views led principals to interpret their role in the evaluation process quite differently. This was true even among principals who shared similar perspectives on the purpose of evaluation, but differed in their views on the best ways to achieve their goals.

Among the principals we spoke with, the vast majority, over 75%, viewed teacher evaluation as a system that should focus on helping teachers improve their practice. This view was shared by principals with a wide range of prior teaching and administrative experience and who led schools at every level. For example, one principal described the purpose as follows:

> I think it's to get feedback to our teachers on the work that they're doing, and how to, number one, how to make sure they know that you're there to support them – but to also let them know where they need support and help, and then help us identify the help that they need to be better teachers.

Many other principals echoed this sentiment stating that, "[the] evaluation process is at its core to improve teacher practice" and that the goal is "to promote learning and growth." This common viewpoint was aligned with the district's messaging of the primary purpose of the new evaluation system.

However, four of the administrators we spoke with explained that they used the evaluation system to support the vast majority of teachers to improve their practice, but also highlighted the importance of dismissing teachers who were ineffective educators. One principal with seven years of experience characterized the dual objective as "to support that teacher to become better. That would be the first goal. The second alternative, not a goal but an alternative, would be to remove that teacher from the profession." This view was most often expressed by

more experienced principals. A principal with five years of experience described the purpose of evaluation as follows:

> It's to improve teacher instruction in order to improve student achievement, to raise student achievement. That's the purpose. If the person isn't meeting a certain standard, then they need to be removed, because we only want the best for our students, only the best teachers in front of our students.

These principals often framed the evaluation system in terms of raising student achievement, a goal that could be accomplished via professional development and the selective dismissal of low-performing teachers. One principal we spoke with even viewed evaluation exclusively as a process for identifying and removing underperforming teachers. She stated plainly, "I think the purpose of evaluations should be to weed out those that aren't doing their job." These different perspectives led to very different approaches to implementing the new evaluation system.

Consequence: Principals used the evaluation process in very different ways. Principals leveraged the evaluation process to achieve a range of goals that were not always aligned or consistent with the district's stated intent. Implementation approaches differed substantially even among the majority of principals who viewed improving teachers' instructional practices as their primary goal of the evaluation process. Some principals emphasized the importance of direct feedback that is "specific and actionable, and that comes from a place of knowledge and experience on the part of the administrator." Other principals saw teacher self-reflection as the primary mechanism for improvement. "I think ultimately the goal is for teachers to self-reflect on their teaching and become better teachers and realize the areas that they need to work on as teachers," stated an elementary school principal with 22 years of classroom experience. One principal who was a veteran middle school teacher focused on a third mechanism - monitoring and accountability - as a means of motivating teachers to improve their practice:

> I believe that administrators better be in the classrooms. That's the only way to improve. Hey, go ahead and drive home today and have a police officer, just by chance, be behind you. You become an infinitely better driver.

These differences in implementation suggest the teachers' experiences with the evaluation process varied considerably across schools, and that principals did not always leverage all possible mechanisms through which evaluation might promote professional growth.

The juxtaposition of two examples helps to illustrate how principals' differing perspectives and individual goals, rather than the district's intentions, determined how evaluation was implemented in schools. One principal we spoke with, an experienced educator and administrator who had been principal at one of the small semi-autonomous high schools in the district for ten years, believed evaluation should only focus on teacher improvement. However, over the years she had developed her own system of observation and feedback cycles that she implemented independently from the evaluation system. She was frustrated that the new reforms now forced her to situate this informal process within the evaluation process. In her view, the complex new system was full of "verbage" and "grandstanding" and led her to adopt a compliance based approach to evaluation that was separate from her informal feedback process.

The veteran principal of a high-performing high school who viewed evaluation as a process for removing ineffective teachers implemented evaluations in ways consistent with her goals. She invested little time evaluating and providing feedback to teachers that met her expectations. Instead, she used the evaluation process to document poor performance and evaluate out low-performing teachers. Teacher development initiatives at her school were focused on a data-driven instruction initiative and collaborations among teacher teams. In both of these schools, the principals were unwilling to use the new evaluation system as a development tool. Both principals saw other approaches such as teachers observing and providing feedback to

23

their peers as more promising avenues for promoting teacher growth. These choices affected the evaluation experience of teachers in their schools, and may even impact the degree to which the culture around teacher evaluation changes in the district as a whole.

Challenge #2: The expanded role of principals. Nearly all principals, 88%, expressed real concerns about the increased demands of the new evaluation system. As one principal put it, "the biggest challenge is time." Principals commonly described the process of evaluating all teachers in their schools as "a nightmare" or "nuts." As one principal shared, "It's too much. It almost killed me to try to do all of it." This view was held by principals of all levels of experience who worked in both smaller and larger schools. The district evaluation plan substantially expanded the role of principals in teacher evaluation without releasing them from any of their other responsibilities. One mid-career elementary school principal likened this experience to sitting down to dinner at a family-style Italian restaurant:

> It's like going to Sorentos. Sorentos is the kind of place where they pride themselves on Italian tradition, right? Educators pride themselves on Italian tradition. That tradition is we're going to keep piling on your plate until it falls over. We're not going to remove anything. If you want to remove something off your plate you'd better eat it. If not, here comes the food. It keeps coming.[3]

Several other principals, including two principals of small elementary schools with few other administrative staff, explained that if they had dedicated themselves fully to the evaluation process "their building [would] fall apart." A principal of a large elementary school asked rhetorically, "What about your buses? What about your cafeteria? What about your parents who want to meet with you? What about your district people who are calling you for this or that?" Unexpected situations required principals to be "out and about, and available." These types of interruptions made it difficult for principals to protect the blocks of time they needed to observe teachers, craft well-written evaluation feedback, and hold post-observation conferences.

Consequence: Feedback conversations were infrequent and brief. The demands on principals and their administrative teams to conduct extensive evaluations for all teachers limited the frequency and quality of feedback teachers' received. Several principals expressed concerns that they were unable to provide the frequent feedback necessary for supporting teachers' professional growth because of the sheer number of teachers they were required to evaluate. From the perspective of one principal, if feedback cycles for improvement are "done right, it's a weekly to monthly thing that you do with teachers." Instead, it was all that most principals could do to observe and write the formative and summative evaluations for each teacher in their school. The high ratio of teachers to evaluators was of particular concern for one principal:

> A leader—or in this case an instructional leader—can only be effective if the feedback and support that they provide is high quality. We know from research in the private sector that a supervisor or manager can only be effective supervising up to 12 people. Once you go beyond 12 people, you're not able to provide the time and attention and support and feedback to those people as you can if you have 12 or fewer… I'm evaluating 48 people. . . I really worry about myself as an instructional leader, because am I really providing quality feedback and quality time and quality supervision to that many people? I personally don't think so.

A principal of a large middle school expressed similar concerns. "In years past I would spend, with maybe a dozen teachers, I would spend a tremendous amount of time. I [would] sort of be very superficial with the rest. This year I was sort of deeper with 40 but not able to get nearly as deep with a few." The infrequent evaluations and limited oversight under the former evaluation system allowed some principals to provide more in-depth feedback to the teachers they felt needed the most support.

Even principals who were able to hold their time dedicated to observations as "sacred" struggled to find time for post-observation conferences. Nearly 90% of our sample mentioned the

limited time available for giving teachers feedback. One principal broke down the time he

dedicated to the evaluation process as follows:

> I would say writing it up is the majority of the time. Evaluation shouldn't be mostly writing, but I think that I would say that it's meeting with teachers that is probably the least amount of time. I'd say that's probably 5-10% of it. Observation is probably 10-15, and then the rest is devoting to writing it.

While the exact breakdown of time varied considerably across principals, this pattern where the

least amount of time was spent on in-person conversations with teachers was quite common.

"The actual face-to-face conversation is not where I wanted it to be," was a common sentiment

expressed by principals with varying levels of experience. This finding is particular concerning

given clear evidence that observation and feedback cycles are most effective when they are

frequent, in-depth, and sustained over long periods of time (Garet, Porter, Desimone, Birman, &

Yoon, 2001; Yoon, Duncan, Lee, Scarloss, & Shapley, 2007).

The responsibility of submitting written evaluation feedback on-line that became part of a

teacher's permanent record also caused principals to shift their focus away from in-person

feedback conversations. The electronic system increased the visibility and permanence of the

write-up compared to the old carbon-copy evaluations that were filed away and often lost in the

paper shuffle. It also served to increase the pressure on principals to draft carefully worded

feedback that balanced accurate assessments with the ability to motivate teachers. An

experienced middle school principal with no teaching experience explained his anxiety:

> I fell into this trap where I would go in and do an observation for 20 min and then it would take me an hour and 20 min to write feedback for the teacher because I was trying to write the perfect piece of feedback where they wouldn't be offended but they would be inspired; where it was authentic and constructive and it wasn't judgmental; where they would follow through on what I was writing in the feedback and they wouldn't just dismiss it as either, "He isn't going to follow-up with me on this," or "I disagree with him." … I was spending no time conferencing with people.

A high school principal echoed these sentiments when she explained that, in an "ideal situation," she would want her written and verbal feedback "to be equal." By closely monitoring the written evaluations, the district created a strong incentive for principals to prioritize formal one-way communication over a more productive two-way dialogue about instructional improvement.

Challenge #3: Providing feedback outside their expertise. Evaluating and providing specific feedback to teachers across subjects and grade levels presented substantial challenges for principals. Nineteen of the twenty-four principals we spoke with expressed concerns about their ability to provide meaningful feedback to teachers in all disciplines and levels. Elementary school principals typically characterized this challenge in terms of grade levels. A principal who taught second grade explained that his "weaker point would be the upper grades." A young principal of a new elementary school explained, "I feel a little bit more comfortable in the upper grades," as he had only taught fifth grade. A third elementary school principal who had also taught fifth grade expressed similar sentiments, "[I] feel a lot more comfortable in grades 2-5. . . The kindergarten world is like a different world."

For middle school and high school principals, evaluating teachers across different subject areas presented a challenge. A principal with five years of experience teaching history and English told us, "history, I do, science and math are a little bit of a challenge." She explained that she preferred to observe math teachers with the math coach. A high school principal laughed at the notion that she was responsible for evaluating foreign language teachers. "What do I know about Spanish and French?" she exclaimed. One middle school principal who taught English language learners for 32 years stated simply, "I am not a math person." Principals often relied on their own teaching experiences as a primary source of ideas for supporting teachers. When they

evaluated teachers in subjects and grades they had not taught, principals felt less comfortable and

confident in their abilities to evaluate instruction accurately or provide meaningful support.

Consequence: Feedback was narrowly focused on pedagogy. Lack of content expertise

led many secondary principals to narrow the focus of their evaluation to general instructional

practices and strategies. Eight principals told us how they focused on pedagogy rather than

content. A veteran high school math teacher who had just become the principal of her high

school explained how she adapted her feedback across subjects. "I just find that, for myself,

whenever I'm evaluating a math teacher, it's very easy to give content suggestions, and I give

pedagogy, but not content [feedback], in the other areas." A high school principal with five years

of experience said that her peers recommend a similar strategy:

> The advice that I got was to really, for content areas that I did not teach, to really focus in
> on just the instruction. To not worry about the content unless there was just something
> egregious.

Another high school principal even went as far as to focus exclusively on pedagogy in the

evaluation process. As she put it, "It's not about the subject. You know what good teaching is

and it doesn't matter what content it is."

One principal we spoke with who had no prior teaching experience approached

evaluation by looking for general practices that he felt were beneficial for students. During

observations he would ask:

> How is the teacher planning to ensure all students are engaged? How is the teacher
> planning to use their time wisely and to be efficient with time? How is the teacher
> planning in terms of differentiating instruction? How is the teacher planning in terms of
> using groups?

This principal also described how teachers at his school had raised the issue of his lack of

content expertise at a faculty meeting. His approach was to be "honest with [teachers]" that they

"are more of experts in each of the content areas than I will ever be." Instead, he explained, he chose to "defer to district experts" when it came to questions about implementing curriculum. Although narrowing the scope of feedback may have improved principal's confidence, it failed to address teachers' need to develop both their core content knowledge and their pedagogical content knowledge, which have been shown to be central elements of effective instruction particularly in math (Wayne & Youngs, 2003; Hill et al., 2008).

Challenge #4: Principals had limited training. The current evaluation system demanded a wide range of skills from principals in order to implement the new process successfully. Principals were required to accurately differentiate teachers on a four point scale, support their ratings with low-inference evidence, communicate these ratings effectively, and prescribe specific, actionable feedback for teachers on how to improve. In the district we studied, evaluator training was focused on familiarizing principals with the expansive rubric and procedural requirements, and calibrating principals to be reliable and accurate raters. At the time, principals had not received any training on how to manage their time to complete all observations or how to engage in productive feedback conversations.

Even with training on how to use the new classroom observation instrument, principals experienced real challenges differentiating among teachers, particularly at the upper and lower ends of the rating scale. Nine out of the 24 principals we interviewed felt the limited training provided by the district was detrimental to implementation, particularly in differentiating levels. A veteran principal of a large elementary school told us, "I think we really have a very, very fine line in between exemplary and proficient." Another experienced administrator described that he and his peers struggled with identifying "the difference between a genuinely bad teacher, who isn't trying to improve, versus a teacher who just doesn't have the skills in place that they need,

and could improve, if they were given the right supports and feedback." The current evaluation system required principals to distinguish between ratings that, in the experience of some principals, required nuanced assessments. Accurate evaluation ratings are critical for any evaluation system, but are also a necessary precursor for engaging in a productive conversation with teachers about professional improvement.

In addition to assigning accurate ratings, there was a critical "human component," as one principal described it, that they had to learn on their own. "It's an area that isn't emphasized," the principal lamented. A high school principal with previous experience as a non-profit manager explained how principals were now expected to know how to teach adults as well as children.

> The way that the role is described, the role of the principal, it says "instructional leader" and you're told to give feedback, but I don't think that there's been a lot of training and resources provided on what that looks like and how to do it well, and how to do it even in challenging difficult relationships.

For principals who transitioned into administration directly from the classroom, the only option was "learning when you get into the job," as one principal explained. These challenges could be even greater for administrators who had no classroom teaching experience. A principal of a large high school with over 100 teachers lamented that "some of our administrators haven't taught, so that's a challenge." These administrators lack of an "instructional lens" meant they gave "very different evaluation responses" than other member of her team. This variability in evaluators' abilities to identify areas of weakness or strength and communicate their feedback had important consequences for the quality of post-observation conversations with teachers.

Consequences: Feedback conversations focused on ratings and positive reinforcement rather than on how teachers could improve. The process of evaluating teachers in a way that supported their professional growth required principals to rate teachers accurately and have

direct conversations about what and how a teacher needed to improve. Differentiating among teachers who had been told they were satisfactory for many years led to feedback conversations that became focused on the summative evaluation rating itself rather than areas for continued professional growth. Rating teachers lower than they felt was fair often derailed efforts to focus the conversation on professional improvement. As one principal described:

> I was pretty communicative and still people would be crying or, 'I can't believe you think that, Needs Improvement, I've never been Needs Improvement.' I wanted to say, 'Well, of course you've never been Needs Improvement, it hasn't existed before.'

A young elementary school principal spoke about how teachers she rated as Needs Improvement would frequently respond, "But I've always met standards." She then had to explain that they "met it barely, minimally," under the old system and that standards were now higher under the current system. Even some teachers who were rated as Proficient were still fixated on why they were not rated as Exemplary rather than on the things they could do to become Exemplary. A veteran high school principal described the situation in her school:

> It creates a lot of tension when you don't label a teacher Exemplary. I mean, I've never had so many people complain about not being Exemplary. It's been more discouraging than encouraging . . . . they feel like they're not appreciated.

While some principals saw these situations as opportunities to talk about how teachers could further enhance their practice, not all principals were prepared to navigate these conversations.

Our interviews also suggested that some principals may have avoided difficult conversations with teachers about their weaknesses and, instead, focused on reinforcing the things that were going well in the classroom. Only three principals we spoke with described how telling teachers that they needed to improve was a challenge for them. However, these principals suggested that many more of their peers would "shy away from difficult conversations." As one administrator described, "The most difficult part of the job is probably to deliver those difficult

messages, and not everyone is capable of that." The focus of the evaluation process on improving teachers' practice meant principals also had to navigate a dual role as supervisor and instructional coach. Another principal explained that her biggest challenge was, "Finding a balance where you say to people, 'I need you to do something really different from what you've been doing. Don't be afraid to make mistakes. Oh, but by the way, I'm your evaluator, so I'm watching what you're doing all the time.'" Decades of research by Anthony Bryk and his colleagues (Bryk & Schneider, 2002; Bryk, Sebring, Allensworth, Luppescu, & Easton, 2010) have demonstrated the key role of relational trust between administrators and teachers engaged in improvement processes. Some principals shied away from using feedback conversations to push teachers on their growth areas for fear of jeopardizing this relational trust.

Assessing Proposals for Improving Observation and Feedback Cycles

A variety of ideas for how to improve the quality of feedback teachers receive emerged from our conversations with principals. Here, we review the most salient proposals, discuss how they relate to the theory of action behind observation and feedback cycles, and assess the degree to which they address the implementation challenges described above.

Reduce the evaluation load. In the district we studied, adding at least two observation and feedback cycles for every teacher to principals' existing responsibilities prevented many principals from dedicating the time necessary to support teachers' with frequent and in-depth feedback. As one middle school principal said, "High quality implementation would've been me working with 12 people." Two different principals suggested they could not work with more than a dozen teachers at a time and be expected to make any real difference in teachers' practices.

Principals could focus on fewer teachers by distributing their evaluation responsibilities more widely among school leadership teams. Several principals of larger schools took this

approach but still struggled to achieve a ratio of one evaluator per every 12 teachers. In at least one instance, it also created challenges when teachers felt that not all administrators were applying the same standards in the evaluation process. Districts could develop a more flexible evaluation system by relaxing annual evaluation requirements or reducing the number of observation and feedback cycles for high-performing teachers. For example, Montgomery County Public Schools require annual evaluations for beginning teachers but experienced teachers are evaluated on 3 to 5 year cycles. Teachers rated as Below Standard are observed and evaluated more frequently and are provided intensive supports. This targeted approach would allow principals to focus their attention on providing frequent feedback to those teachers who were most in need of improvement. However, several principals we spoke with also warned of the risks associated with this approach. Teachers may perceive evaluations as a process for collecting evidence to justify dismissals and be unwilling to openly recognize their weakness and engage in the improvement process. Requiring all teachers to participate equally in a rigorous evaluation process sends a strong signal that the process is not exclusively focused on dismissal.

Shift operational responsibilities. A second potential solution to principals' limited time which came up in nine different interviews was to narrow their responsibilities to focus primarily on instructional leadership. Principals commonly described instances when their investments in instructional leadership were undercut by unexpected operational issues or constrained by their other building responsibilities. One principal lamented:

> We spend a lot of time doing a lot of operations work, following up on phone calls, following up on emails; time, and time, and time again, which pulls us away from the classroom, or having conversations with teachers.

Several principals saw these operational responsibilities as directly limiting their evaluation practices. "My whole job could be evaluation, easily, but I also have to run a building,"

explained a principal at a combined middle and high school. A middle school principal proposed that, "If they want the principal to be an instructional leader, taking as much of the operations out of their purview as possible is probably what needs to happen."

Restructuring principals' roles to focus less on operations management could serve to substantially expand their capacity to provide evaluation feedback. Several charter school networks such as Uncommon Schools and Success Academy Public Schools have adopted formal co-leadership models with instructional leadership and operations management positions (Frumkin, 2003). Moving towards more task specialization among administrators is promising given the increasing demands on principals to be expert instructional leaders and the core importance of operations management.

Train and coach principals. A third common proposal we heard from over half of the principals we interviewed was to provide targeted training and coaching for principals. This could involve efforts to increase principals' time management skills as well as their ability to use observation and feedback cycles to drive instructional improvement. Some principals thought the district could do more by, for example, "providing more models of how to structure a regular meeting with teachers [and] how to lay out your calendar effectively." A veteran teacher and principal stated, "Ideally, we should be getting feedback about our feedback." A younger principal of a large middle school echoed these sentiments:

> I'm always interested to do a better job at providing people feedback ... The 'Good job, keep it up,' feedback doesn't go very far, you know? You want be more specific about teaching and teaching strategies that you can give to them.

Principals recognized that they were being asked to develop and deliver feedback in a way that was new and more demanding than many had experience with.

Providing better training to principals is an intuitive solution, but little is known about the content and efficacy of such training programs (Peterson, 2002). The Wallace Foundation's National School Administration Manager (SAM) program helps principals reallocate time from managerial tasks to instructional leadership by documenting their time use, identifying areas for greater efficiency, and training principals how to build staff capacity to manage operations and respond to common situations independently. SAM staff also provide coaching to improve principals' instructional leadership capacity. Evaluations of SAM programs found that principals gained nearly an hour a day to focus on instruction (Turnball et al., 2009), but small to no effects on student achievement (Turnball, White, and Arcaira, 2010). Maximizing principals' time management skills and ability to distribute tasks can help them meet the increased demands for their instructional leadership, but research shows that dedicating more time to instructional leadership may not be sufficient to promote teacher development and student achievement (Grissom et al., 2013). Only high-quality training on how to conduct observation and feedback cycles would address principals' limited expertise. Even high-quality training focused on the feedback process is unlikely to change the practices of principals that are focused on removing ineffective teachers or to address a lack of content knowledge.

Hire instructional and content experts to coach teachers. One veteran principal we spoke with found that the demands of the new evaluation system meant that he "could not spend a lot of time coaching." He described how instead, he hired full-time instructional coaches to work closely with his teachers. Several principals saw the need for coaches who were content experts to supplement the general instructional feedback they could provide. "I'm advocating that the district actually put together a network of content leaders . . . Let's have them also take some responsibility in evaluating depth and knowledge of content," said a veteran high school

principal. Similarly, another principal told us, "let's have some direct evaluation of real understanding of content by people who are district-wide specialists."

A growing body of research suggests that instructional and content coaches can improve teachers' practice through sustained observation and feedback cycles (e.g. Allen et al., 2011; Authors, in press). A system where coaches work across schools would allow districts to better match teachers to experts in their particular area in need of improvement. Content experts would also be well prepared to rate and provide teachers feedback based on content-specific observation rubrics such as the MQI or the PLATO (Kane et al., 2014). However, coaching models require substantial financial investments to sustain the high frequency of coaching cycles found to be most effective (Authors, in press). Without a dedicated financial commitment to coaching, this approach might simply replace one implementation constraint (principals' time constrains) with another (the high cost of individualized coaching).

Develop peer observation and feedback systems. Twelve principals we spoke with suggested that peer observation and feedback systems held more promise for promoting professional growth than relying on principals to provide evaluation feedback. Many of these principals emphasized the value of providing teachers with opportunities to observe and learn from their peers. As a principal at one large high school described, "I think the best way to improve instruction is to put together a system where teachers actually go in and observe each other." An elementary school principal explained how peer-to-peer observations system are "a great opportunity for teachers to see other teaching styles, other teaching techniques and to really realize that they can improve their own teaching with the staff that's right there."

The principals we spoke with framed peer observation as a method for improving instruction outside of the evaluation process. The literature on distributed leadership provides

examples of how principal could empower their staff to assume responsibilities for instructional leadership and development (Camburn, Rowan, & Taylor, 2003; Leithwood et al., 2007; Spillane, Halverson, & Diamond, 2001). For example, new evidence documents the potential of pairing highly effect teachers to work with their less effective colleagues on specific areas of instructional improvement (Papay, Taylor, Tyler & Laski, 2016). Dozens of districts have also adopted peer evaluation systems where expert teachers assume formal responsibility for evaluating their peers (Papay & Johnson, 2012). Peer Assistance and Review (PAR) is one of several examples of how districts can enable expert teachers to conduct rigorous observations and provide detailed feedback that supports professional growth. PAR can increase teachers' impact on students achievement (Taylor & Tyler, 2012) and can be cost effective (Papay & Johnson, 2012), but requires effective labor-management cooperation as Hillsborough County Public School's decision to scrap its established peer evaluation system illustrates (Sokol, 2015).

Conclusion

Over a quarter century ago, Popham (1988) wrote about the "dysfunctional marriage" of formative and summative teacher evaluations. In his view, evaluation systems can help teachers become more effective, or dismiss inept teachers from their positions, but not both. Today, teacher evaluation systems are undergoing sweeping changes in order to increase their rigor and reliability for high-stakes decisions, as well as to provide teachers with actionable feedback to support improvement. It remains an open question whether these reforms are capable of reconciling the marriage of teacher development and dismissal in one single system.

In the large urban district we studied, reforms to the teacher evaluation system provided a common framework and language that aided principals in assessing and discussing teachers' professional practice. Principals' perceived that teachers were becoming more involved in the

evaluation process and that the culture around evaluation was beginning to shift towards a focus on professional growth. They described how teacher buy-in and investment in the improvement process were essential to its success. These changes provided necessary structures and more fertile contexts for principals to promote growth among their staff as evaluators. However, the expanded role of principals as evaluators resulted in a variety of unintended consequences.

Principals described a variety of challenges associated with implementing observation and feedback cycles that limited their ability to promote teacher development. Differing perceptions about the purpose of evaluation among principals, teachers, and the district sometimes undercut the trust and buy-in required for meaningful conversations about instructional improvement. Pushing all teachers to recognize and address their own areas for improvement after being rated satisfactory for many years made for challenging conversations. Many principals also described how the expanded demands to observe all teachers multiple times constrained the quality and depth of feedback they could provide. Expectations to provide detailed feedback to teachers outside of principals' grade-level and content-area expertise resulted in a focus on content-free pedagogical practices. Finally, the district's focus on compliance caused principals to prioritize written feedback over in-person conversations. These unintended consequences illustrate that *how* an evaluation system is implemented ultimately determines whether it will be successful at promoting teacher development.

While our interviews provide a window into the implementation challenges principals can face as evaluators, this case study has several limitations. Our study captured a snapshot of principals' experiences in one district at a single point in time. Principals' perspectives will vary depending on the design of the evaluation systems adopted in their districts and the specific stage of implementation. The district we studied had not yet incorporated measures based on students'

standardized test scores into its evaluation system. Furthermore, principals were responsible for

assigning overall evaluation ratings whereas in most states summative ratings are calculated from

a weighted sum of multiple performance measures. These differences limit the generalizability of

our findings across different contexts. Lastly, our small sample of 24 principals limited our

ability to analyze potential differences across school contexts in the degree to which principals

were successful at supporting teacher development through evaluation. Recent research suggests

that future studies should be designed to specifically examine differences in how principals

promote professional development across school contexts (Authors, 2014).

Our assessment of the potential solutions to the implementation challenges principals

faced as evaluators point to several avenues for addressing unintended consequences. The

absence of evaluator training programs focused on the feedback process is a major

implementation barrier (Herlihy et al., 2014). An effective training and support program for

evaluators could help them to better manage their time and maximize the impact of their

evaluation feedback. Our findings also highlight the increasing need to develop principals' skills

as evaluators and instructional leaders as part of their graduate training and certification

programs. However, no amount of preparation and training will resolve the challenges related to

principals' lack of experience with some subjects and grades or their time constraints.

Consolidating operations management responsibilities into one primary administrative position

to allow principals to focus on instructional leadership is one possible solution. Another would

be to spread evaluation responsibilities between principals and peer evaluators. This is a

particularly promising approach given the emerging evidence for these models, the ability to

match teachers with peer evaluators that have relevant content and grade-level expertise, and the

potential to integrate the peer evaluator position into a broader career ladder system for teachers.

The remaking of teacher evaluation systems across U.S. public schools has the potential to promote teacher improvement on a large scale. Delivering on this promise will depend, in large part, on how these reforms are implemented on the ground by administrators and educators.

Endnotes

1. We do not provide references for these sources in order to maintain the confidentiality of the district.

2. We retrieved information on the teacher evaluation rubrics and evaluation procedures from each district's website.

3. The name of the restaurant is a pseudonym

References

Almy, S. (2011). *Fair to everyone: Building the balanced teacher evaluations that educators and students deserve*. Washington, DC: Education Trust.

Authors (2012).

Authors (2014).

Authors (2015).

Authors (in press).

Boyd, D., Grossman, P., Ing, M., Lankford, H., Loeb, S., & Wyckoff, J. (2011). The influence of school administrators on teacher retention decisions. *American Educational Research Journal*, *48*, 303-333.

Bryk, A., & Schneider, B. (2002). *Trust in schools: A core resource for improvement*. New York, NY: Russell Sage Foundation.

Bryk, A., Sebring, P. B., Allensworth, E., Luppescu, S., & Easton, J. (2010). *Organizing schools for improvement: Lessons from Chicago*. Chicago, IL: The University of Chicago Press.

Camburn, E., Rowan, B., & Taylor, J. E. (2003). Distributed leadership in schools: The case of elementary schools adopting comprehensive school reform models. *Educational Evaluation and Policy, 25*, 347-373.

Center on Great Teachers and Leaders. (2014). *National picture: A different view*. Retrieved March 31, 2014 from http://www.gtlcenter.org/sites/default/files/42states.pdf.

Center on Great Teachers and Leaders. (2013). *Databases on state teacher and principal evaluation policies*. Retrieved September 1, 2015 from http://resource.tqsource.org/stateevaldb/.

Cubberly, E. P. (1916). *Public School Administration*. Cambridge, MA: The Riverside Press

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

Danielson, C., & McGreal, T. L. (2000). *Teacher evaluation to enhance professional practice*. Alexandria, VA: Association for Supervision and Curriculum Development.

Danielson, C. (2007) *Enhancing Professional Practice: A framework for teaching.* Alexandria, VA: Association for Supervision and Curriculum Development.

Donaldson, M. L. (2012). *Teachers' perspectives on evaluation reform*. Washington, DC: Center for American Progress.

Donaldson, M. L. (2013). Principals' approaches to cultivating teacher effectiveness: Constraints and opportunities in hiring, assigning, evaluating, and developing teachers. *Education Administration Quarterly, 49*, 838-882.

Donaldson, M. L., & Papay, J. P. (2014). Teacher evaluation for accountability and development. In H. F. Ladd & M. E. Goertz (Eds.), *Handbook of research in education finance and policy*. New York: Routledge.

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.

Frumkin, P. (2003) *Creating new schools: The strategic management of charter schools*. Baltimore, MD: Annie E. Casey Foundation.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes

professional development effective? Results from a national sample of teachers. *American Educational Research Journal, 38*, 915-945.

Grissom, J. A., Loeb, S., & Master, B. (2013). Effective instructional time use for school leaders: Longitudinal evidence from observations of principals. *Educational Researcher, 42*, 433-444.

Hallinger, P., Bickman, L., & Davis, K. (1996). School context, principal leadership, and student reading achievement. *The Elementary School Journal, 96*, 527-549.

Halverson, R. R., & Clifford, M. A. (2006). Evaluation in the wild: A distributed cognition perspective on teacher assessment. *Educational Administration Quarterly, 42*, 578-619.

Halverson, R., Kelley, C., & Kimball, S. (2004). Implementing teacher evaluation systems: How principals make sense of complex artifacts to shape local instructional practice. *Educational administration, policy, and reform: Research and measurement*, 153-188.

Hanushek, E. (2009). Teacher deselection. In D. Goldhaber & J. Hannaway (Eds.), *Creating a new teaching profession* (165-180). Washington, DC: Urban Institute Press.

Herlihy, C., Karger, E., Pollard, C., Hill, H. C., Kraft, M. A., Williams, M., & Howard, S. (2014). State and local efforts to investigate the validity and reliability of scores from teacher evaluation systems. *Teachers College Record*, *116*(1), 1-28.

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction, 26*, 430-511

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.

Honig, M. I. (Ed.) (2006). *New directions in education policy implementation: Confronting complexity.* Albany, New York: State University of New York Press.

Horng, E. L., Klasik, D., & Loeb, S. (2010). Principal's time use and school effectiveness. *American Journal of Education*, *116*, 491-523.

Hoy, A.W., & Hoy, W. K. (2012). *Instructional leadership: A research based guide to learning in schools*. 4th Edition. Pearson.

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

Kane, T., Kerr, K., & Pianta, R. (2014). Designing teacher evaluation systems: New guidance from the measures of effective teaching project. John Wiley & Sons.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment.* MET Project. Seattle: Bill & Melinda Gates Foundation.

Kennedy, M. M. (2005). *Inside teaching: How classroom life undermines reform*. Cambridge, MA: Harvard University Press.

Ladd, H. (2011). Teachers' perceptions of their working conditions: How predictive of planned and actual teacher movement? *Educational Evaluation and Policy Analysis, 33*, 235-261.

Lavigne, A. L., & Good, T. L. (2015). Improving teaching through observation and feedback: Going beyond state and federal mandates. New York: Routledge.

Lacireno-Paquet, N., Morgan, C., & Mello, D. (2014). *How states use student learning objectives in teacher evaluation systems: A review of statewide websites.* Washington, DC: U.S. Department of Education.

Leithwood, K., & Louis, K. S. (2011). *Linking leadership to student learning*. San Francisco: John Wiley & Sons.

Leithwood, K., Mascall, B., Strauss, T., Sacks, R., Memon, N., & Yashkina, A. (2007) Distributing leadership to make schools smarter: Taking the ego out of the system. *Leadership and Policy in Schools, 6*, 37-67.

Louis, K., Dretzke, B., & Wahlstrom, K. (2010). How does leadership affect student achievement? Results from a national US survey. *School effectiveness and school improvement*, *21*, 315-336.

Maxwell, J. A. (2005). *Qualitative research design: An interactive approach.* Thousand Oaks, CA: SAGE Publications.

May, H., & Supovitz, J. A. (2011). The scope of principal efforts to improve instruction. *Educational Administration Quarterly, 47*, 332-352.

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

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

Newmann, F. M., Smith, B., Allensworth, E., & Bryk, A. S. (2001) Instructional program coherence: What it is and why it should guide school improvement policy. *Education Evaluation and Policy Analysis, 23*, 297-321.

Patton, M. Q. (2001*). Qualitative research and evaluation methods*. (2nd Ed.). Thousand Oaks, CA: Sage Publishing.

Papay, J. (2012). Refocusing the debate: Assessing the purposes and tools of teacher evaluation. *Harvard Educational Review, 82*, 123-141.

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

Papay, J., Taylor, E. S., Tyler, J., & Laski, M. (2016). Learning Job Skills from Colleagues at Work: Evidence from a Field Experiment Using Teacher Performance Data. NBER Working Paper No. 21986

Peterson, K. (2002). The professional development of principals: Innovations and opportunities. *Educational Administration Quarterly*, *38*, 213-232.

Popham, W. J. (1988). The dysfunctional marriage of formative and summative teacher evaluation. *Journal of Personnel Evaluation in Education*, *1*, 269-273.

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

Robinson, Lloyd, Rowe 2008

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

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

Sebastian, J., & Allensworth, E. (2012). The influence of principal leadership on classroom instruction and student learning: A study of mediated pathways to learning. *Educational Administration Quarterly, 48*, 626-663.

Spillane, J. P., Halverson, R., & Diamond, J. B. (2001). Investigating school leadership practice: A distributed perspective. *Educational Researcher, 30*, 23-28

Spillane, J. P., & Kenney, A. W. (2012). School administration in a changing education sector: The US experience. *Journal of Educational Administration*, *50*, 541-561.

Sokol, M. (2015, October 29). Hillsborough schools to dismantle Gates-funded system that costs millions to develop. *Tampa Bay Times*.

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

Steinberg, M. P., & Sartain, L. (in press). Does teacher evaluation improve school performance? Experimental evidence from Chicago's excellence in teaching project. *Journal of Policy Analysis and Management*.

Strauss, J., & Corbin, A. (1998). *Basics of qualitative research: Grounded theory procedures and techniques*. (2nd Ed.). Thousand Oaks, CA: SAGE Publications.

Stronge, J. H. (2005). *Evaluating teaching: A guide to current thinking and best practice*. Corwin Press.

Supovitz, J., Sirinides, P., & May, H. (2010). How principals and peers influence teaching and learning. *Educational Administration Quarterly, 46*, 31-56.

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

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

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

Tucker, P. D. (1997). Lake Wobegon: Where all teachers are competent (or, have we come to terms with the problem of incompetent teachers?). *Journal of Personnel Evaluation in Education*, *11*, 103-126.

Turnball, B. J., Haslam, M. B., Arcaira, E. R., Riley, D. L., Sinclair, B., & Coleman, S. (2009). *Evaluation of the School Administration Manager Project*. Washington, DC: Policy Studies Associates, Inc.

Turnball, B. J., White, R. N., & Arcaira, E. R. (2010). *Achievement trends in schools with School Administration Managers (SAMs)*. Washington, DC: Policy Studies Associates, Inc.

Waters, T., Marzano, R. J., & McNulty, B. (2003). *Balanced leadership: What 30 years of research tells us about the effect of leadership on student achievement*. Aurora, CO: Mid-continent Research for Education and Learning.

Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research, 73*, 89-122.

Weisberg, D., Sexton, S., Mulhern, J., Keeling, D., Schunck, J., Palcisco, A., & Morgan, K. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. Brooklyn: The New Teacher Project.

Witziers, B., Bosker, R. J., & Kruger, M. L. (2003). Educational leadership and student achievement: The elusive search for an association. *Educational Administration Quarterly, 39,* 398-425.

Yoon, K. S., Duncan, T., Lee, S. W. Y., Scarloss, B., & Shapley, K. (2007). *Reviewing the*

evidence on how teacher professional development affects student achievement. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest.

Table 1
Principal and School Demographic Information

	Interviewed	Non-Interviewed	*p*-value
Principals Characteristics			
African American	0.46	0.39	0.54
Caucasian	0.38	0.44	0.60
Hispanic	0.08	0.16	0.32
Asian American	0.08	0.01	0.06
Male	0.42	0.28	0.21
Age (years)	47.52	47.21	0.90
School Characteristics			
Elementary	0.46	0.41	0.66
Middle	0.13	0.06	0.27
High	0.17	0.21	0.65
Traditional	0.63	0.69	0.58
African American (%)	34.76	34.75	1.00
Hispanic (%)	41.47	44.46	0.48
White (%)	11.54	12.46	0.76
Asian (%)	10.05	5.52	0.06
Independent Education Plans (%)	17.03	19.12	0.18
English Language Learners (%)	29.00	29.55	0.89
Low Income (%)	70.06	71.02	0.77
Proficient in English language arts (%)	49.29	46.99	0.64
Proficient in mathematics (%)	42.57	41.80	0.86
Observations	24	86	

Note: P-values are derived from two-sample *t*-tests of the mean difference in a given characteristic across interviewed and non-interviewed principals. Proportions of schools that are elementary, middle, and high school do not sum to one because of schools with non-traditional grade configurations.
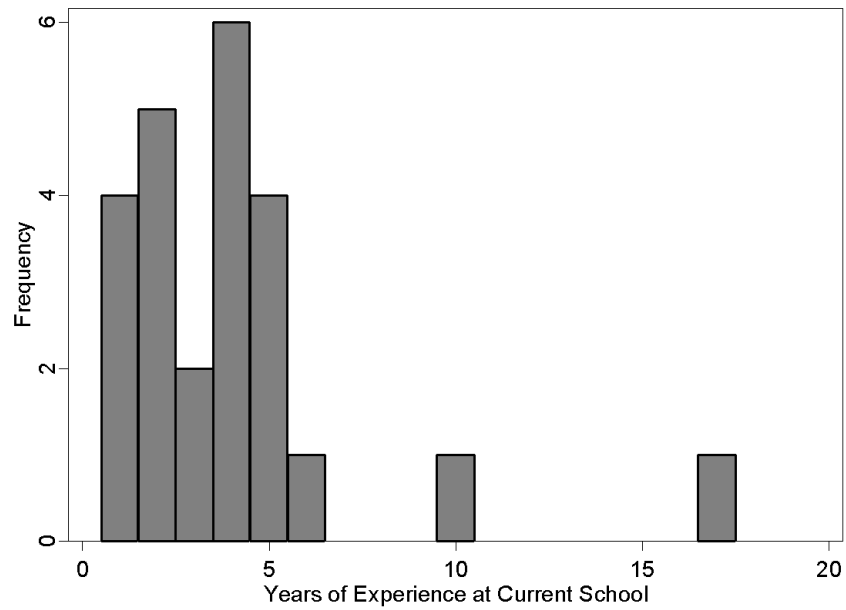
Figures

Panel A
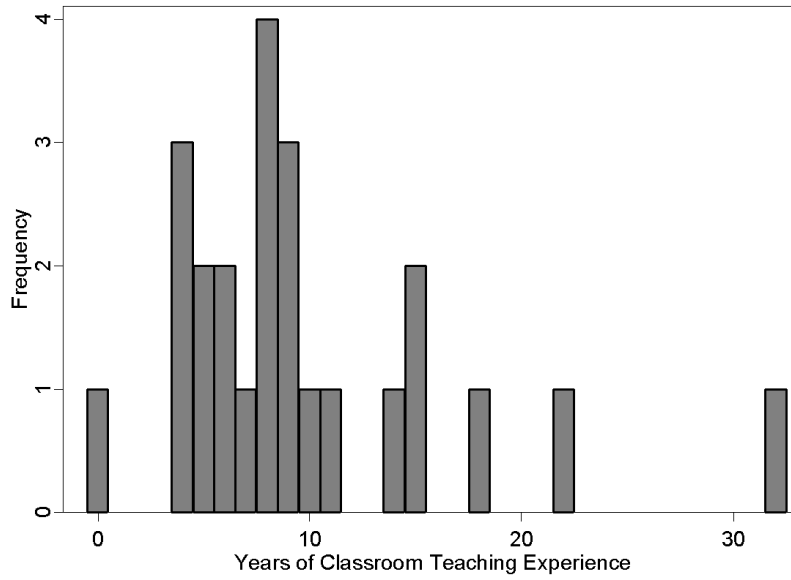
Panel B

Figure 1: Histograms depicting distributions of the total number of years of classroom teaching experience and total number of years of administrative experience at current schools for interviewed principals.