Evaluation, Pay for Performance, and Teacher Learning around Instruction:

Inducements, Incentives, and Obligated Practices

in the Organizational Life of Public Charter Schools

by

Rick Mintrop
Miguel Órdenes
Erin Coghlan
Laura Pryor
Cristóbal Madero

University of California, Berkeley

Paper Presented at the Annual Conference of the American Educational Research Association, Chicago, 2015

Since the mid-1990s, with the advent of the accountability era in some pioneering states, policy makers in the U.S. have experimented with new forms of public management in education that incorporated principles from the private sector. In so doing, they followed trends in public management that have been vigorously discussed in the public administration literature under the auspices of 'reinventing government' (Osborne & Gaebler, 1992) or 'New Public Management' (Aucoin, 1990; Ferlie, Ashburner, Fitzgerald, & Pettigrew, 1996; Hood, 1991). Measurement through standardized indicators, goal setting, performance monitoring, quasi-markets, and incentives have been the hallmark of these reforms. For education, the *No Child Left Behind* law made this approach pervasive across the United States. Under its auspices, organizational units, such as schools and districts, were held accountable to *outcomes;* and in case of missed performance targets, a staged sequence of sanctions applied to the organization. Laws that authorize charter schools facilitated quasi-market conditions.

Beginning in 2008 with the American Recovery and Reinvestment Act (ARRA) the federal government signaled a management approach that went beyond organizational accountability for outcomes. The federal Teacher Incentive Fund initiative (TIF)[1] drilled down to individual teacher performance by way of value-added test results and evaluations of teaching with presumed consequences for careers and benefits. TIF (Center for Educator Compensation Reform, 2014) stipulated that recipients of bonus pay grants, such as public school districts, charter management or support organizations, use a mix of indicators to identify educators at various levels of performance and reward or sanction them differentially via merit bonuses or promotions. Performance was supposed to be calculated based on standardized tests, where possible on the 'value added' of individual teachers, scores on rigorous teaching evaluations that

[1] Source: http://www2.ed.gov/programs/teacherincentive/index.html

were externally scored, principal ratings of teachers' progress, and participation in school development activities. Thus TIF required the schools to adopt an extensive performance measurement system for teachers and administrators that was displayed in an intricate data dashboard consisting of multiple measures and material rewards (Center for Educator Compensation Reform, 2014).

This paper reports on a study of three public charter schools in the state of California that adopted the TIF design as its performance management system. Over a period of four years, we traced the development of beliefs, attitudes, motivations, practices, and results among teachers and administrators in response to this management system. Our intent is not to evaluate whether TIF worked according to program intent. There are several rigorous evaluation studies available in the literature (Glazerman et al., 2011; Watson et al., 2010) . Nor do we want to relate an implementation story that would document predictable shortcomings in clarity, validity, or capacity, given the complexity of the system (Taut & Sun, 2014). Instead, we want to answer a more pointed question: how did the presence of a TIF-inspired performance management system in the lives of the schools influence teacher learning around instruction?

When schools or school systems adopted TIF, they added the program onto the existing state accountability framework that was largely test-outcome and sanctions-driven along NCLB regulations (Mintrop & Sunderman, 2009). The schools on which this study reports had been accustomed to this accountability framework for about 13 years at the time of adopting the new TIF incentive structure. By standardizing and externally evaluating instruction, attaching extrinsic monetary rewards to these evaluations, and creating explicit performance statuses for effective and ineffective teachers, TIF adds a new dimension to performance management that potentially moves key improvement levers from outcomes and assessments back to inputs. A

paramount purpose of TIF is achieving better student outcomes through higher quality instruction. Higher quality instruction may result when teachers focus their learning on instruction, use tools and artifacts for analyzing and judging the planning and execution of lessons or lesson segments, and engage in evidence-based and analytical conversations with colleagues and instructional supervisors. The interplay of teaching evaluations, bonus payments paid according to evaluation scores, and the use of tools and artifacts, obligated by TIF is the focus of the study. Practices and artifacts obligated by TIF are formative classroom visits by instructional supervisors, a video of a lesson submitted to external evaluation and rated according to an observation instrument, called the Summative Evaluation of Teaching (SET), and bonus pay. Will the new evaluative criteria, the new bonus, and the novelty of lesson videos move teacher learning forward? Will it make learning about instruction more central, more precise, and more results oriented?

A Conducive Environment for Performance Management

Some scholars have argued that evaluation and pay for performance cannot develop their full potential because of *political* factors (Podgursky & Springer, 2006). When it comes to evaluation and pay for performance, it is often governments or higher level administrators that pursue management reforms while the lower level performance units are expected to implement them (Pollitt & Bouckaert, 2011). High-stakes accountability systems were introduced this way (Debray, 2006; Shipps, 2006). Public sector institutional contexts, such as public school districts, have shown political resistance to management reforms along private sector lines (Trujillo, 2013). Unions are known to defend single salary scales and tenure (Podgursky, 2006), tenured teachers eviscerate evaluations (Milanowski & Heneman, 2011) and administrators

accommodating teachers' micro-political power may shy away from necessary conflict (Timperley & Robinson, 1998).We chose to study cases in which these types of political dynamics would presumably play less of a role, allowing us to observe performance management dynamics without interference from district and union politics. We did so deliberately in order to test the assumption that a deregulated labor regime (Barzelay, 2001) would be a more opportune environment for the positive stimulus of performance management.

Our cases are public charter schools that volunteered to take on TIF as a new performance management system. TIF was adopted by school-level administrators at the initiative of a local charter support provider. The leaders of these schools may not have known the concrete shape of the system, but they could presumably see the main architecture.

Teachers in these schools were non-tenured, non-unionized, and were paid on a salary schedule that allowed for differential pay beyond seniority. Contract renewal from year to year depended on performance and was in the discretion of school administrators and governing boards. Not atypically for these types of schools, the three schools, but especially two of them, tended to recruit from a pool of young novice teachers with uncertain commitment to stay (Ingersoll, 2001; Johnson, 2007). As independent non-networked schools, the schools experienced the full weight of accountability, managerial independence, and market competition. They existed under the watchful eye of the local school district with which they competed for students and which could make the re-authorization of their charters difficult. In other words, unlike typical public schools, the three charter schools existed in a marketized and deregulated space, commonly thought to be conducive for performance management.

We want to know how under these circumstances, external teaching evaluations and pay for performance, together or separately, shape how teachers internally engaged in learning

around instruction. Towards this end, we first discuss relevant literature on teacher evaluation, pay for performance in education, and teacher learning around instruction. These literatures provide theoretical concepts and conjectures that guide data collection and analysis. Next we describe the study's detailed longitudinal data set and the ways with which we analyzed it. The report of findings is organized along three distinct time periods over four years that begins with the adoption of TIF. For each time period we look at the three main dimensions of the study, evaluation, pay for performance, and teacher learning and their interactions with each other.

Relevant Literatures

Although the name suggests that the Teacher Incentive Fund, as an experimental policy instrument proffered by the U.S. federal government, is mainly about incentives, for the jurisdictions that are involved in the program, TIF is at the heart an inducement. Often times the terms inducement and incentive are treated as synonyms, but for our purposes a clear distinction is necessary. Inducements are transfers of money to agencies in return for the production of certain goods that the government values. Inducements often come with regulations spelling out focal activities that recipients are expected to carry out. Hence they obligate specific practices (McDonnell & Elmore, 1987, pp. 138–139). For example TIF spells out that monetary bonuses need to be disbursed contingent on performance in multiple dimensions and that teaching evaluations should be "objective measures" that require a rigorous evaluation and teaching observation tool (Max et al., 2014). An incentive, by contrast, has the connotation of motivating an outcome or a performance, to kindle or encourage hard work or effort as workers anticipate rewards (Mitchell, Ortiz, & Mitchell, 1987). TIF brings resources to participating jurisdictions in return for implementing practices that locally may, or may not, produce performance-contingent

incentives. Employees may implement the required practices in return for money, but whether they feel compelled to increase effort and improve practices according to stipulated evaluative indicators is a different matter. In an inducement logic, schools may comply with obligated practices in return for money without actually incentivizing performance.

Evaluation

Teacher evaluations are, and have been, a widespread and common practice in educational systems. In the last fifteen years, in the wake of standards-based reform, the face of teacher evaluation has changed. Traditionally, evaluations of teachers were conducted by way of administrator observation often based on unreliable instruments that produced unstructured, global judgment (Stodolsky, 1990). As a result, teacher evaluations did little to drive improved teacher effectiveness and increased student achievement (Toch & Rothman, 2008). Newer standards-based systems, such as those using the MET instruments, seek to reform traditional teacher evaluations through judging teachers with the help of more rigorous instruments (Danielson, 2011; Pianta, Paro, & Hamre, 2005) and augmenting observational measures with additional measures, such as value-added scores on student achievement tests as well as student survey results (Bill and Melinda Gates Foundation, 2013).

Evaluations have a summative and formative purpose. Summatively, evaluations are to assure quality by creating fixed performance statuses based on teacher effectiveness measures that can potentially undergird accountability and supervisors' decisions about promotions or dismissals. Formatively, evaluations render diagnostics and feedback that help employees learn, grow, and improve. School instructional leaders, especially those with authority over subordinate employees, play a double role. For summative purposes, they are supervisors who make performance statuses transparent and communicate the stakes (Blasé & Kirby, 2008; Blase &

Blase, 1998; Goldstein, 2007). For formative purposes, they are facilitators who provide expert feedback(Glickman, 2002; Kluger & DiNisi, 2006), motivate improvement efforts with support and encouragement, and channel feedback into teacher collegial learning and capacity building (Louis, Marks, & Kruse, 1996). Formative and summative purposes and aspects of instructional leadership associated with them are difficult to integrate with one another (Darling-Hammond, 2013; Darling-Hammond, Wise, & Pease, 1983; Millman & Darling-Hammond, 1990). However, current evaluation reform initiatives attempt to combine formative and summative approaches for the purposes of both teacher professional development and high-stakes salary and personnel decisions (Bill and Melinda Gates Foundation, 2013; The New Teacher Project, 2009).

The literature on instructional leadership and supervision indicates that instructional leaders have been neither strong supervisors willing to make high-stakes decisions based on evaluations (Darling-Hammond et al., 1983; Murphy, Hallinger, & Heck, 2013; Tucker, 1997), nor expert diagnosticians providing powerful feedback (Darling-Hammond, 2013; A. Milanowski & Heneman, 2011; Scriven, 1974). The literature is rife with complaints that evaluation and instructional supervision have played a largely symbolic role in the lives of schools (Darling-Hammond, 2013; Hill, Kapitula, & Umlad, 2010; Kane, Kerr, & Pianta, 2014). In the instructional leadership literature, a view prevails that accentuates leaders facilitating organizational conditions for commitment to teacher learning around instruction in a collegial manner that respects teachers' classroom autonomy (Bryk, Sebring, Allensworth, Easton, & Luppescu, 2010; Knapp & Feldman, 2012). According to this literature, it makes sense for instructional supervisors to emphasize regard and collegiality over evaluation and sanctions. Recent education reformers, however, assert that teaching and learning will not improve without high-quality feedback based on accurate teacher assessments measured against clear, research-

based standards (Archer, Kerr, & Pianta, 2014). An evaluation system grounded in precise and accurate standards allows administrators to create actionable teacher improvement plans as well as informed personnel decisions; these two components are essential for raising student achievement (Archer et al., 2014).

For teachers, the new evaluative measures may implicate the self in varied ways and depending on how teachers feel implicated, they may process the messages of the system in a more expedient superficial or thorough and systematic fashion (Gregoire, 2003). In the NCLB-like accountability systems, a summative performance index based on standardized tests is calculated for the organization as a whole. These sorts of measures are only loosely linked to teacher practices and work effort and are often hard to interpret. Thus, they provide only weak guidance and feedback (Hatry, Greiner, & Ashford, 1994; Supovitz, 2002), and only weakly implicate teachers as competent actors (Finnigan & Gross, 2007; Mintrop, 2004), though they may spur effort on the part of teachers to prevent negative repercussions for the organization. Information from student assessments may be linked to practice in more fine-grained ways (Raudenbush, 2004; Hill et al., 2010). Direct evaluations of teaching practices, however, once lifted out of triviality and perfunctory procedure may stimulate strong self-implication among teachers. Such evaluations may become an overt marker of performance and competence or a signal that may evoke evaluative threat, status anxiety, and stress (Donaldson, Gooler, & Scriven, 2002; Sansone & Harackiewicz, 2000). Teachers may distance themselves from the evaluations and exhibit defensiveness or detachment (Gregoire, 2003) or they may engage with these evaluations as affordances for new learning.

Research on standards-based evaluations, implemented over the last few years (Hatry et al., n.d.; Kimball, 2002; Milanowski & Heneman III, 2001; Odden, Kelley, Heneman, &

Milanowski, 2001) shows that defensiveness or detachment may be avoided when teachers value the evaluation criteria as an expression of their own internal standards of teaching excellence, when they believe that the evaluation system renders judgments that are procedurally and distributively fair, when the feedback is precise enough so that teachers know how to improve, when they believe they have the capacity or efficacy to reach satisfactory performance (Kimball, 2002; Milanowski & Heneman III, 2001), and when they consider the evaluator credible and trustworthy (Stiggins & Duke, 1988).

In a performance management system, such as the one adopted by the three schools, teaching evaluations are both formative and conducted internally as well as summative and conducted externally. Internalizing external evaluative judgments, for example a rating or score on a video of a lesson, may be challenging given the potential divisiveness of performance status (Buunk, 2003) and traditional accommodation between evaluating supervisors and evaluated teachers in schools (Darling-Hammond, 2013; Millman & Darling-Hammond, 1990; Murphy et al., 2013). How could external summative criteria be made fruitful for internal teacher learning? Administrators could use their authority and leverage evaluations to communicate a shared framework of good teaching to which they expect teachers to perform and/ or be held accountable (Bill and Melinda Gates Foundation, 2013; Kane et al., 2014). Instructional coaches and other instructional supervisors could use summative criteria, observation instruments, and other artifacts related to the performance management system when they give formative feedback on teaching. Teacher leaders or teams of teachers working together in learning communities could adopt external evaluation criteria as guidance and internal cues of competence. Performance statuses would, thus, become internalized as a symbolic marker of professional excellence (Milanowski, Heneman, & Kimball, 2011; Sansone & Harackiewicz,

2000). Alternatively, organizational leaders, coaches, and teachers together, may buffer dynamics of external evaluative judgment through silencing, ignoring, trivializing, not participating, de-valuing, etc. , in the process relegating external evaluative judgments to the periphery of organizational life. When this happens, evaluate judgments may still be present, tolerated as a necessary quid pro quo for bonus monies, but they become latent features of organizational life with perhaps muted motivational force. However, they may still loom large individually, depending on individual performance status.

Pay for Performance

In public education, pay for performance (PfP) is a management approach that tries to uproot the tradition of a single salary scale by creating visible performance criteria and applying them transparently so that educators continuously receive monetary rewards throughout their careers for above average performances, most notably student achievement and teaching excellence (Podgursky & Springer, 2007). Those who advocate PfP hypothesize that if a performance measure can be agreed upon and teachers are rewarded on that metric, teacher motivation and performances valued by the organization, the educational system, or even society would align (Lazear, 2003). Rewards would motivate teachers to work harder and encourage organizational dynamics to be more goal driven (Podgursky & Springer, 2007; Yuan et al., 2012). PfP schemes follow an expectancy model of motivation (Vroom, 1964; Lawler, 1973) that pivots on extrinsic rewards valued by workers and the belief that one has the wherewithal to attain the reward with one's effort.

The pay for performance literature, in summary, tells us that PfP schemes have had a patchy track record in education as to their impact in student outcomes, effects on beliefs and practices, and sustainability over time, both during the 1980's (Murnane & Cohen, 1985) and

nowadays (National Research Council, 2011; Springer, 2009; Yuan et al., 2012). The literature is less strong on telling us why this is the case. Some recent studies have shed some light on this phenomenon (Glazerman, at al. 2012; Marsh et al., 2011; Springer et al., 2011, 2012). They find that pay for performance is often implemented with serious flaws as to reliability of procedures, validity of measures and payouts, required knowledge, skills, and capacity, and meaningfulness for practice (Yuan et al., 2012). PfP schemes seem to be more effective when they engage educators in activities that enhance capacity (Glazerman, McKie, Carey, & Harris 2012) although it is not clear from this literature whether the positive effects are due to the monetary aspects of the scheme or the practices that it obligate educators to engage in (Taut, 2007).

The literature on the use of pay for performance in other industries suggests that PfP works best when employees are largely extrinsically motivated (Lazear, 2000), for example in sales, where the use of bonuses are overt features, performance statuses are sharply distinguished based on clear metrics, and performance status exerts a forceful motivational punch (e.g. Banker et al., 1996). When employees are largely intrinsically motivated, fulfill complex tasks that are difficult to capture with simple metrics, or work in teams, bonuses become a double-edged sword (Frey, Homberg, & Osterloh, 2013). They may create unproductive competition and demoralization by crowding out employees' intrinsic motives and commitments to work teams. To prevent this from happening, lower-level managers and subordinate employees tend to deemphasize, background, or relegate bonuses to the periphery of organizational life even when they are mandated by upper management. Bonuses may become latent, fuzzy, and perhaps muffled in their punch (Beer et al., 2004). The bonus may still be there, alluring individuals, but as a marker of overt reward for good performance they are weakened. In this context, the finding from some recent evaluation studies (Glazerman et al., 2012; Marsh et al., 2011;

Springer et al., 2011, 2012) may make sense that majorities of participating teachers did not consider pay for performance harmful for collaboration or unfair. The prevailing sentiment seems to have been one of embracing the chance of making extra money (e.g. Marsh et al., 2011). Alternatively, bonuses, as extrinsic rewards, may become symbolic markers that cue competence (Harackiewicz & Sansone, 2000) when they are linked to intrinsically meaningful performance metrics. In our case, meaningfulness of bonuses for teacher learning and instructional effort would depend the meaningfulness of evaluation metrics.

Obligated Practices

Acceptance of grant monies from the federal TIF program obligates schools to create specific learning tools or "artifacts" that engage teachers and leaders in specific practices that are regulated by the grant. Regardless of the dynamics related to rewards or evaluative judgments, artifacts have the potential to develop a life of their own and shape how organizational members (teachers, coaches, instructional supervisors) think about their work and how they structure learning and improvement (Halverson, Kelley, & Kimball, 2004). The infiltration of artifacts influences the language with which individuals process their cognitions (Cole & Engstrom, 1993). While the potential of artifacts may be reinforced through the combined impact of rewards and evaluative judgment, the literature on distributed cognition has shown how to trace and uncover the ways artifacts by themselves seep into existing practices of organizations and transform them (Gery, 1991; Halverson & Clifford, 2006; Hutchins, 1995; Nemeth, O'Connor, Klock, & Cook, 2006; Solomon, 1993).

As mentioned, the concept of artifacts derives from distributed cognition, a theory that posits that cognition does not take place in an individual's mind alone, but is a distributed

phenomenon that takes into account historical conditions, culture of a group, social interactions, and artifacts within a given environment (Hutchins, 1995). In the case of organizations, artifacts are the tools, routines, policies, or symbols that serve to reduce complexity and structure work behavior (Hutchins, 2006). Within organizations, artifacts give testimony to the cognitive system of the organization, and the functional-system interaction between artifacts and cognitive processing is shared across organizational members (Hutchins, 2006). Relevant artifacts for the purposes of the TIF project are the "Samples of Effective Teaching" (SET) tool, the production of lesson videos, the tools used, and protocols followed, in formative feedback on teaching based on the California Standards of the Teaching Profession), the TIF data dashboard, and protocols followed in teacher inquiry groups or professional development sessions.

Roughly speaking, the distributed cognition literature explores the interaction between artifacts and their users by looking at the constraints and affordances that artifacts engender (Halverson et al., 2004; Hutchins, 2006; John & Sutherland, 2005; Nemeth et al., 2006). In the case of formative and summative evaluations, a lesson observation tool affords an analytical lens through which the complexity of teaching can be captured, but it also constrains this lens by way of the observational behavioral indicators that the tool privileges, alongside the evaluative threat the tool may pose. A video of a lesson creates visibility of ordinarily shielded instructional practices, but a summative stand-alone video also constrains what is seen. As actors (or users) interact with these tools, the tools or artifacts inscribe themselves individually and collectively into actors' cognitions, but the artifacts are made sense of in the context of existing beliefs, understandings, knowledge, skills, norms, values, and practices that are shared within the organization (Hutchins, 1995). Artifacts are thus translated and incorporated into these cultural contexts whereby individuals or groups adopt, ignore, or modify the artifact, or eliminate it if it

does not fit into the distributed cognition of the group members (Hutchins, 1995). The process of successfully integrating artifacts into existing organizational cultures of schools (Halverson, 2003; Halverson et al., 2004) can help to reshape attention, make goals more precise, make performances more visible, and in this way enhance organizational learning (Hutchins, 2006; Nemeth, Cook, O'Connor, & Klock, 2004). When artifacts and their attendant practices are translated into tacit beliefs and taken for granted routines, they regulate what the organization does (Hutchins, 1995).

At the most basic level, TIF obligates teachers to engage with specific artifacts in return for resources that accrue to the organization as a whole. These obligations may be communicated by authoritative leaders who want to ensure bonus monies through compliance with regulations. TIF artifacts may inscribe themselves into the schools if they are seen as resources for learning and improving instructional practices. As organization-wide tools, they may span cross-functional roles and may generate advocacy by those who discover new uses or affordances for learning.

The literatures on teacher evaluation, pay for performance, and the role of artifacts in organizational learning generated theory-guided codes that were used as first and second level coding schemes for qualitative data.

Methods and Data

The data for this study are part of a longitudinal study that began in the 2011-12 school year. Data from three school years are analyzed. The school year 2011-12 was the first year of implementation, the school year 2013-14 was the last school year before the sun-setting of the project. We collected data in three secondary schools that we treated as three separate cases. But

soon we discovered that a good portion of the response patterns, especially related to evaluation and pay for performance, applied across all three schools. Learning patterns differed to a larger degree depending on established learning routines and cultures at the schools. In the findings section, we report on overall patterns and point out where patterns differed from school to school.

The local TIF project grant was directed by a nonprofit organization. This nonprofit supported the implementation of the TIF grant at the three schools. The director of the nonprofit as well as the principal of School A (at the time of the grant) conceptualized and wrote the TIF grant. Once awarded, the TIF evaluation system was designed by the nonprofit director and school leaders from all three schools. Additionally, the evaluation design incorporated feedback from teacher focus groups. The local system consisted of multiple student assessment measures, formative and summative teacher evaluation, and participation in leadership and collegial learning activities. The overall number of measures was 26.

This paper examines processes around two scores: the Formative Evaluation of Teaching (FET) and the Summative Evaluation of Teaching (SET). The FET was conducted as a somewhat informal classroom visit and conversation between supervisor and teacher, aided by the extensive and broadly constructed California Standards of the Teaching Profession. The SET was a clinical tool that was used formally once per year. Teachers submitted a video of their best example of teaching to external evaluators. The video would be scored on a scale of 1-4, a rating of "3" qualifying for a bonus. The rating could account for a $1,000 bonus per year. Schools had access to the SET tool year-round.

The SET observation instrument was developed according to the schools' specification. They opted for a relatively basic lesson design format, consisting of six components: activating

prior knowledge and interest, introducing new content through modeling, co-constructing dialogue, or student exploration, checking for understanding and guided practice, independent practice, feedback, and synthesis. For each component, behavioral indicators directed teachers to pay attention to student time on task, active participation, and academic engagement. Principals were also awarded performance bonuses as a result of the quality of their instructional conferences with teachers, judged by an observation tool developed by the local TIF provider.

Cases

The three schools are relatively small in size (see Table 1). Two of the schools are high schools, one combines middle and high school grades. The schools are urban in character. The schools are located in distinctly poor sections of a metropolitan area in northern California. Within the state accountability system in the school year 2011-2012, School A performed strongly for its demographic profile with an Academic Performance Index (API) close to 800 (the state's target). Schools B and C, on the other hand, were poor performers by state standards.

Table 1. Demographics and Performance 2012

	School A	School B	School C
API 2011 Base	567	C	787
API 2012 Growth	554	542	758
Enrollment	176	85	178
Black or African American (%)	4.0%	12.9%	13.5%
Hispanic or Latino (%)	55.7%	72.9%	80.3%
White (%)	9.1%	5.9%	1.7%
Two or More Races (%)	15.9%	1.2%	0.6%
Socioeconomically Disadvantaged (%)	58.5%	89.4%	87.1%
English Learners (%)	84.7%	69.4%	40.4%
Students with Disabilities (%)	5.1%	4.7%	11.8%
Note: C means the school had significant demographic changes and will not have any growth or target information.			

The schools serve a challenging population of economically marginalized students of color and immigrant background. The three schools have in common that they blend charter school autonomy with a strong avowed commitment to social justice and serving their marginalized communities.

Data

A mix of qualitative and quantitative data were collected. This paper only addresses the qualitative data. We conducted 8 rounds of semi-structured interviews. All in all, 52 teachers and 15 administrators participated in data collection over a period of three years. A total number of 130 interviews were carried out, augmented by 65 hours of observations of faculty meetings and meetings at grade and subject matter levels during which learning around instruction took place.

Table 2. Summary of Interviews and Meeting observations per school

	School year	School A	School B	School C	Total
Rounds of teachers interviews					
1. Base line	2010-2011	4	3	6	13
2. Follow-up implementation.	2011-2012	2	1	5	8
3. Teacher learning.	2011-2012	3	-	6	9
4. Bonus release 2011-12 and teacher motivation.	2012-2013	4	8	7	19
5. Teacher learning follow-up	2012-2013	1	-	6	7
6. Bonus release 2012-13	2012-2013	4	4	4	12
7. Exit interview year	2013-2014	9	6	8	24
8. Bonus release 2013-14	2013-2014	3	4	6	13
Total teachers interviews	**-**	**30**	**26**	**48**	**105**
Rounds of leaders interviews					
1. Administrators Baseline interview	2010-2011	2	1	2	5
2. Administrators Follow-up interview	2012-2013	1	2	2	5
3. Administrators Final interview	2013-2014	2	2	3	7
1. Instructional leaders Baseline interview	2010-2011	0	0	1	1
2. Instructional leaders Follow-up interview	2012-2013	0	0	0	0
3. Instructional leaders Final interview	2013-2014	2	0	5	7
Total leaders interviews	**-**	**7**	**5**	**13**	**25**
Rounds of observations					
Meeting observations (hours of observation)	2013-2014	15	10	40	65

Codes for analysis of interviews were developed following the concepts derived from the relevant literatures (Miles, Huberman, & Saldaña, 2013). The interviews were coded using a qualitative data analysis software (Dedoose). We developed 47 distinct codes, the majority of them derived from theory. Some additional codes captured emergent phenomena. The theory-derived codes were defined, operationalized, and illustrated with representative quotes from interviews. We trained a team of four coders for inter-rater reliability. Twenty percent of interview excerpts were coded by two coders and the discrepancies were treated collaboratively

in order to clarify the concepts among the coders. The codes were grouped in five conceptual complexes: evaluation of teaching, performance-contingent payment bonus, teacher learning, school context, and leadership. Here we display a selection of main codes for each complex:

Table 3. Main Code for Each Complex

Complex/Codes	**Definitions** **Excerpts that refer to…..**
Performance-contingent bonus payments	
Bonus: performance contingency	…bonus payments, considered (or discarded) as reward for special effort or alternatively as generic inducement to reward good work. [specific number or effort (the criteria)…]
Bonus: acceptability	…bonus payments as (non)acceptable/important/valuable way to reward work effort, quality, or competence.
Bonus: validity	…understanding of the mechanics of the allocation of bonus monies (how bonuses are calculated and disbursed) and the associated judgment of validity and fairness.
Bonus: expectation of reward	…the belief that effort and/or learning will likely result in desired bonus payments.
Bonus: usefulness for practice	…reports on increased engagement with activities *deliberately* targeted towards attaining the bonus. Especial attention should be put on instruction.
Bonus as latent allure	…inferences that implicitly hint at money playing a muted role in teachers' calculations. Bonuses become de-coupled from performance through silencing , ignoring, trivializing.
Bonus: negative consequences	…reports on any observed negative repercussions from bonus payments, such as demoralization (crowding-out), distortion of practices, unproductive competition.
Evaluation of Teaching	
Eval: Self-reflection	…teachers self-reflecting on their own teaching and judging their own quality of teaching/ competence / self-efficacy without clear feedback from others…
Eval: Feedback to teaching	…formal communication received from formative or summative evaluation of teaching…
Eval: FET	… perceptions of FET: clarity, usefulness, fairness, precision of judgment, goal setting.
Eval: SET	…perceptions of SET: clarity, fairness, transparency of external evaluation, usefulness, precision of judgment, goal setting
Eval: self-implication	… respondents treating evaluations with attention (or inattention) due to: - administrator/authoritative expectations, - colleagues' expectations, - consonance with ones' criteria of good teaching, - dissonance between evaluative expectations and one's own performance Or the lack thereof.
Eval: threat, stress appraisal	… appraisal of capacity of self, work team, or school as a whole to meet self-adopted goals, school (IL's, coaches, colleagues) expectations.
Eval: depth of engagement	…teachers' dispositions to connect to evaluations (especially FET/SET) with resistance, subversion, detachment, non-participation, expediency, learning engagement.
Eval: as latent concern	…inferences that hint at evaluation judgments or scores playing an implicit/ muted / non-overt role in teachers' self assessment. They become de-coupled

	from performance through silencing, ignoring, trivializing, but are still present as a "bug."
Teacher learning	
Collaborative professional learning	….collaborative learning that occurs in the organization. Individual teachers participate in professional learning activities within the school (ex: inquiry groups, grade-level meetings, PD, coaching, etc.)
Professional learning individual	….individually initiated professional learning that results in teacher participating in professional learning outside of his or her school. These can include relationships with other teachers not at the school (i.e. seeking out Master's programs, collaboration with teachers at other schools, etc.)
Feedback	...communication received from those exposed to one's work (students, parents, colleagues, supervisors) about the quality of one's teaching that is useable as inputs to inform one's teaching (i.e. classroom observation, students comments, tests scores, etc.)
Norms of learning together	…descriptions of how teachers ordinarily relate to each other when learning together (i.e. norms of equality, making differences in expertise visible, assuming self-determination, openness or closeness of the tasks, etc.)
Video: affordance / constraint	…any mention of using video to film teaching and provide feedback; this includes SET videos.
SET/FET: affordance / constraint	…. the SET/FET providing orientation for improving teacher's work, collaborative work in inquiry groups or grade-level meetings, or being incorporated into leadership routines (such as instructional rounds and training new teachers).
Leadership	
Strategies for instructional improvement	…how leaders foster strategies for teacher learning, whether via developing relationships, performance systems, lesson study, professional development, etc.
Artifacts	…precision of diagnostics, goal setting, and feedback associated with learning artifacts (ex: videos, formative and summative evaluations)
Goals	…benchmarks leaders set for growth in teacher instruction and organizational improvement
Adoption of PBECS	...any time leaders describe why their school adopted PBECS in the first place, what they desired from the system, or their reflections on the money associated with the grant.
Organizational culture	…how leaders describe the organizational culture of their school, whether they emphasize relationships or academics, and how they respond to external versus internal policies. Also includes general descriptions of the organizational structure.
Evaluation of PBECS	…how leaders perceive the evaluation component of the FET and SET, and how the usefulness of the data dashboard, the student achievement metrics, and the video artifact.
Problem awareness/growth needs	…where leaders identify growth needs, especially around teaching quality
Vision philosophy	…leaders' general outlook on the school, and whether the vision is team-oriented, community-oriented, or learning-oriented.
School context	
Routines/norms	…formal work structure, stable organizational arrangements, established ways of conducting work and habits, norms, expectations, etc. Examples include inquiry groups, instructional rounds, coaching cycles, and grade-level meetings.
Student descriptions	…any sociological facts, such as demographics, socio-economic status,

	neighborhood, and community.
Micro-politics	…discretion, control, hierarchy, locus of decision-making power, and authority.
Relationships	…relationships among stakeholders of the schools (teachers, students, administrators, parents
Mission	…the broad future aspirations for the school, the organizational philosophy, and the central guiding ideas of the purpose of the organization.

The coding proceeded in four steps: (1) We divided the extensive interview material among the team and coded according to the four main broad descriptors: evaluation, bonus pay, teacher learning, and school context. Two more broad descriptors were coded, motivations and goal-setting, that are not analyzed for this paper. (2) Material under each broad descriptor became the specialty of initially two coders and, once inter-rater reliability was established, one coder. The two coders stayed in close touch to discuss uncertainties in coding. All interviews under each broad descriptor were coded with theory-guided sub-codes and a few emergent codes. (3) Once all interviews were coded in this way, we sorted the coded interviews according to the rounds during which the interviews were conducted. Each round took place at a particular time. The "data dashboard release," i.e. the release of performance data and bonus payments each year, were critical incidents that marked stages in the life of TIF. These stages roughly followed Year 1, 2, and 3 of implementation. (4) We compared and looked for patterned differences across time for each code, in the process reducing data with the help of a meta-matrix (Miles et al., 2013) that summarizes patterns in concrete descriptive language for each relevant code and each stage.

Findings

We untangle dynamics related to evaluation, rewards, and induced practices obligated by tools or artifacts, and we trace their interplay over three years. In a nutshell, in the three charter schools the adoption period was characterized by *consonance*, the belief that the TIF performance management system could be implemented with relative ease and that costs of

implementation outweighed the benefits. In mid-life, *dissonance* set in. Performance contingencies attached to both bonus and external evaluations were perceived as disconfirming the values of the schools. They were largely rebuffed and relegated to the periphery. Once the power of incentives became latent, a period of *resonance* set in. During this phase, the high point of the TIF performance management system, first administrators and then teachers came to interact with the two main artifacts, videos of lessons and a clinically oriented observation tool, in novel ways. We report findings in three sections that roughly follow the three years of implementation.

Consonance

Adoption of the TIF project began when the principals and the nonprofit support provider (the "provider") received the TIF grant. From the beginning this group of leaders entertained two motives: to garner additional resources for the schools and teachers, and to implement a performance management model that had the potential to improve teacher instruction and student learning. The Executive Director of the provider organization articulates these dual goals best:

> "And initially, I looked at it, and I said, "You know, I'm not convinced about paying for test scores, and I'm just not certain that this is up our alley." And she [school leader] said to me, "You're thinking about this all wrong. We don't have to just submit a proposal that's a simple test scores go up, you get bonuses. We can really use this as an opportunity to really think more broadly about what do we want to incentivize? And how can this be a tool and a lever for improving schools?" So really, a combination of just wanting to get more resources to schools, especially get more resources to teachers, and the opportunity to try to think a little bit differently about how we could design this, is what enticed me to do it."

While the TIF money was clearly a strong motive, the management philosophy behind TIF, its stress on evaluation and its intent to improve instruction by rewarding high performing teachers and leaders made sense to the adopting leaders. The early affirmation and excitement supporting the evaluation component of TIF is captured by an instructional leader at School C:

> My goals for the evaluation system or what I hope their goal is, is sort of twofold. One is to streamline what we think a School C teacher is and what School C practices are. This is our tenth year. We have to be able to say to new staff, here's what School C teachers do. Here's what we're looking for…And I think the second goal of the evaluation should be consistency of

> practice. And by that, I don't mean that every teacher should be teaching the same way because that's crazy. But instead, that we're meeting some sort of expectation, that all students are engaged, that our content is rigorous. (School C, 125)

Data, evidence, evaluation, incentives, and rewards, the buzzwords of new performance management system, held a certain appeal and did not seem to be in conflict with the strong collegial cultures that these leaders had established in their schools and held dear. As described by an instructional coach at School C:

> On the whole, everybody at School C is willing to work together and help each other out and has collegiality, truly. And part of the reason I'm excited about having the evaluation system is so we can say here's what we want. And here's what we don't want. Because I think sometimes it's hard for [our principal] to say that – to just say this isn't working. (School C, 208)

The leaders' philosophy was strongly echoed in the sentiments of the vast majority of teachers. Educators in all three schools felt that they were called to serve challenging student populations with social-emotional and learning needs beyond those of typical students. Addressing these needs, leaders and teachers advocated for a school mission that placed the quality of relationships among students and between students and adults in the center. Working on strong social-emotional bonds was seen as a prerequisite of students' academic success. Below are quotes from each school exemplifying this sentiment:

> I think most people in our school work really hard. They spend a lot of time caring for the students…I think the faculty at this school is politically motivated in the sense that this is really personal for them. This is not a stepping-stone to another career but these are people who want to be teachers, went through a credential program, sought out the school and find themselves aligned with the vision of the school and the outcomes for the students. (School A, 423)

> I think that it's so nice to be somewhere where I feel like I trust everybody, and everyone cares about the students. I don't work with anybody where I'm, like, oh, what are they really doing this for? I know that everybody has sort of a shared vision of how much we care about doing the best thing for the students…I think we're serving a really important population of kids that aren't served in lots of other places. (School B, 304)

> We've defined five tenants of what we do…and I think we've kept really good fidelity to it. We have the rigorous curriculum, keeping high expectations for all students, a component of knowing each student and family well, and meeting the needs of those people and those groups on an individual basis—really making that a discipline that we have and treating each member of our community as an individual—and then there's the component of knowing families well and the last one which is I think the most descriptive—which is knowing, valuing, and trusting teachers as professionals. (School C, 232)

The three schools varied in the degree to which they stressed academics. Common was the idea that the students' marginal social status in society should not deter them from going to college and that it was the role of teachers to work hard for this goal and to further social justice in society with their unstinting engagement and pro-social commitment to students. One of the schools had a reputation for being strongly focused on academics and it had the above-average test scores for its demographic to show for it. One school connected its focus on relationships with the academic goals of cultivating critical thinkers and critical citizens. At the outset, a strong service and social justice orientation towards marginalized students did not seem to conflict with the TIF management philosophy of measuring performance and rewarding educationally differentially.

Established Adult Performance and Learning Culture

All three faculties functioned with the widely shared assumption that everybody at the schools worked very hard and did indeed go the extra mile. Teachers and administrators held that additional incentives to increase effort were actually *not* needed. Given faculty attrition, especially high in Schools A and B, but not as much of a problem in School C, principals aimed at cultivating their longer-term staff to stay at the school longer than three years. To foster teacher growth, leaders at each school provided support, encouragement, and recognition for longevity. Each year, a few teachers, deemed ineffective by the administrators, were not asked to return, but their numbers were very small. For administrators in Schools A and B, getting teachers to stay was a looming challenge in the face of which 'weeding out' ineffective teachers paled as a concern. In all schools, salary incentives were used to reward effective senior teachers and those teaching in hard-to-staff subjects.

Instructional supervision was handled in the spirit of support and openness. Principals reported that they tended to observe their teachers informally and then debrief with the observed teachers over email or through brief conversations in person. The California Standards for the Teaching Profession (CSTP) were used as a broadly structured framework to generate a metric of accomplishment. But it was used in a casual way. The most important objective of supervision was to socialize teachers into the culture of the schools. As a result, direct feedback on instruction would be blended with suggestions about collegial learning and fitting in with the "ways of the school." One principal stated that his purpose of using formative feedback was to communicate the "School C-way" of doing things.

When the project started, teachers at schools A and B reported that external professional development was occurring infrequently. Like some others, a teacher in School A described professional development this way:

> There's a little bit of PD, like we've targeted certain things – ELL, socio-emotional relationship building, so those are our two things that surfaced in the early spring…that was informative for those, and I think on a figuring out level, that would be more systematized. That makes sense. But it wasn't articulated. It was like, "Oh, that makes sense to do this," so there wasn't a whole lot of transparency, not with the intent of trying to be non-transparent, but we're figuring it out as we go. (School A, 422)

In School B, teachers learned together as a whole faculty once per week, and the meetings were focused on building social emotional relationships between teachers and students. Administrators at the school would also meet with teachers informally throughout the year to provide coaching and feedback on instruction. Once a year, administrators would meet with teachers during a conference to discuss teachers' performance management plan (PMP). Administrators provided more formal feedback to teachers during the PMP conversation and defined goals for professional development. The principal described the PMP this way:

> I mean, it's clear but it's not very effective for teachers because there's only a small section about teaching and it's too generalized. And so a lot of it has to do with, like, other – like, sort of other kinds – other responsibilities like calling home to parents, which is all part of teaching, but I'm

> talking about, like, implementing curriculum and instructions with helping in the classroom. There's very little if anything about that. (School B, 317)

School C had an established approach to professional learning. All teachers regularly met in inquiry groups as grade-level or subject matter teams, and beginning teachers were served by a group of experienced instructional coaches with subject matter expertise. The inquiry groups met to discuss student work and student behavior, collaborated on lesson planning, and strategized for longer-term goals, such as getting all students to graduate on-time and apply to college. School C was further along in their model to coach teachers than the other schools, but the observation, coaching, and feedback structure was loose and informal. An instructional coach at School C drives home this point:

> We haven't received anything like a rubric or what it [an evaluation rubric] will look like in the future. My teachers – we all have room to grow. I certainly do, maybe more than others. I spent a lot of last year being like, okay, is this what you're looking for? Here is what we're doing. I've never in my life of teaching ever received a formal, real evaluation of more than four minutes…This year, [the observations] received in our classrooms, I would say pretty much were the same amount as last year. But we received more feedback. It's not like great or on a scale. It is simply here's what I was looking for. Here's what I saw. Here are questions that arise…So for me, I would like different, more rigorous observations. And through inquiry group, we can make that. Okay, cool, all the kids are engaged. They're gonna do it, no matter what. So here's the next level of growth…But being able to set observation goals in inquiry group and be observed on those by other teachers teaching the same concept and then by the director, helps push my practice more. (School C, 208)

The adoption of the TIF performance management system unfolded within these loosely structured and collegial adult learning environments, with teachers and leaders at the time of adoption desiring more specific feedback with defined structures and routines to improve teaching and student learning.

Evaluation of Teaching: The Tool

Concurrent with adoption, a system for evaluating teachers had to be crafted. TIF provider, school administrators, and university-based researchers collaborated to design an observation tool. Several teacher focus groups gave input. Early on, the idea held sway that a

relatively simple tool that aimed at basic effectiveness of lessons and that allowed for precise feedback was needed for purposes of evaluation. Even though this idea conflicted with the strongly critical and constructivist teaching philosophy that permeated the schools, both provider and administrators opted for this approach – as a start.

The Sample of Effective Teaching (SET) instrument, following in the tradition of clinical supervision, was developed out of these guidelines. It broke down lessons into main teaching functions, such as *activating prior knowledge and interest, introducing new content through co-constructing dialogue, checking for understanding and guided practice, independent practice,* or *giving feedback,* and it attached low-inference behavioral indicators to these functions. To give a flavor of the observation instrument, for *activating prior knowledge and interest,* the tool would list the following teacher moves and student cues:

Table 5. Excerpt from Sample of Effective Teaching Instrument

Teacher moves	Student cues
Calls students to attention	Pay attention to teacher
Communicates lesson objectives, topics and/or expectations in language understandable for sts	
Stimulates curiosity and creates 'hook' for new content	Exhibit interest in new content
Recalls prior knowledge and/or experience	Connect to prior knowledge and/or experience
Generates questions or hunches about new content	Brainstorm relevant questions, hypotheses, or hunches related to new content
	Participate widely

Similar lists of teacher moves and students cues were developed for other lesson components. The tool was to be introduced in a series of four professional development workshops over the first year that were to:

- Introduce the instrument and embed it into other metrics schools were already using for teaching evaluations;
- Use the instrument to analyze lessons or lesson segments;
- Practice self-ratings of videoed lesson segments
- Debrief ratings and feedback from external evaluators to submitted video lessons.

Evaluation of Teaching: Teachers' Voices

Teachers appreciated that they received formative feedback in a supportive and constructive spirit:

> I respect the people who evaluate me, and their opinions. (School A, 403)

> It's nice to have another pair of eyes and someone who is far more experienced than I am to help me grow and also just the way that it's done like in a very loving way that is honest. (School B, 309)

Yet, more was demanded. The wish was voiced frequently that supervisors, coaches, or mentors would look at teaching more closely, that they actually "would encourage [teachers] to change [their] teaching practice (School A, 403)." A teacher imagined the wish that:

> ….there was a skill that I identified with my mentor teacher that I wanted to improve on and then there was an assessment based on that improvement. (403, School A).

Other teachers wished for feedback beyond what they were currently receiving:

> I felt that he hasn't been in my classroom enough or he has, but we haven't had a dialogue about it, so you kind of get that, "I'm not really sure where I stand in terms of where his expectations are for me and where I actually am." (School C, 222)

> She actually tells me how many students were engaged, and made that, you know, made me realize oh wow, like, so now I'm more aware of how students I leave out, you know? Two out of fifteen are not engaged, right? But I wasn't given the tools as to how to engage them. (School B, 319)

More specifically, many teachers indicated that they would like more precision when receiving feedback:

> But what I would hope for it would be just like qualitative feedback from experts about what I was doing well and what could be improved. Maybe just some actionable steps for moving forward in my teaching. (School C, 237)

> I think that having the system by which we were being monitored and given the feedback and helped to improve, not just saying, "Here's where you're screwing up," but having you as a teacher identify, "This is something I am struggling with," and then hope your mentor says, "Okay, let's look at this and figure out what we can do about it. Let's figure out what we can try." (School A, 409)

> I've never gotten feedback that said, "I saw you doing this. These are my questions about it. Next time I think you should do this." Usually the feedback I have gotten has been more in the form of like, "I notice. I wonder. I appreciate. And my questions are" something like that. (School C, 222)

Evaluation of Teaching: View from the Outside

The program evaluators conducted independent observations of lessons across the three schools using the SET instruments during the early adoption phase to establish a baseline of instructional quality in the three schools. These data confirmed the sentiment of the concerned teachers. While a few teachers taught exemplary lessons, for the most part, lessons would have benefited from close clinical attention:

- a large percentage of lessons during which observers could not ascertain what was new content and what was review for students;
- many practice lessons with little new stimulating content, opening phases, so called *do-now's*, that drag out for half a period;
- new content being introduced through modeling and teacher monologue; little explicit feedback; few instances of students learning from error or misconception, and
- very few lesson closures other than "exit tickets."

Table Six displays the mean for each lesson component during fall 2011 for a sample of 29 teachers.

Table 6: Mean SET Lesson Components Fall 2011

Lesson Component	Fall 2011 Mean
Engaging	2.05
Co-constructing/ Modeling	1.61
Checking for Understanding	1.27
Independent Practice	2.04
Feedback	1.49
Synthesis / Closure	1.26

Ratings from 1-4, 3 'applying', inter-rater reliability of 75 percent

These data were shared with the provider and the principals of the three schools.

Response to External Feedback on Lessons

In School C, the response was swift. The principal set off a wave of concern for lesson closures. While it would not be correct to say that teachers pervasively engaged with the SET,

many of those who did had a positive impression of the evaluation instrument. Some teachers found value in the structured approach to teaching lessons. They focused on lesson openings and closings while meeting in their collaborative inquiry groups:

> I think it has impacted what we're doing in inquiry group and that has impacted my instructions, so it's a little indirect I think. I think because we shaped Inquiry Group around the outcomes we needed, I think we … we did a whole practice cycle in the Fall and then in the Spring it was like the second time we were doing it for the SET and so I think – and I feel like that was really valuable. We did a really good job planning that lesson and then the end result was – so we did that twice. And that whole process was really good. (217, School C)

At School C, some of the inquiry groups began using the SET tool and the different components, such as openings and closings, to help backwards plan their lessons. Teachers also applauded that because of TIF, their administrators were required to be in classrooms more frequently and engage in formative evaluation and quarterly conferences. The idea had been that quarterly formative feedback was to lead to a summative performance evaluated with the SET. This would have required instructional supervisors to use the SET as a clinical tool, but this did not take place. Rather, formative feedback was, as it had been in the past, informed by the broad standards of the CSTP that placed little constraints on supervisors' choices for conversation topics.

Thus, TIF-inspired teacher evaluations seemed to be off to a good enough start. There was a need for clinical supervision and learning, at least recognized by some vocal teachers. Teachers clamored for more precision in feedback and were eager to learn. Performance evaluation per se was not considered detrimental to established learning cultures, and visibility of instruction was prized over privacy. But evaluation was associated with precise feedback and learning, not primarily with judgment or reward.

Bonus money

Extra money is always welcome was the pervasive answer at the beginning of TIF. Most teachers considered monetary rewards as recognition and validation for effort and commitment already

expended on their work, and not as an incentive to motivate more effort or new behaviors. Money was an inducement to keep doing what one had been doing all along; and this inducement generated positive feelings. Comments abounded, such as these: *It is nice to be recognized.-- I am thankful for the money. Money is a way to compensate us for the hard work we are doing.* Administrators and the provider framed the TIF project in this way as well. For the hard work that teachers were doing, they often stated publicly, they deserved higher salaries than what the schools could pay. And TIF was an opportunity for the schools to attract more funds to augment salaries. Educators agreed with their administrators that rewards were deserved. And TIF money was a token recognition of this universal deservingness. Here is the voice of a teacher working in the school that was deemed very successful by the criteria of the state accountability system:

> This whole process has been very interesting for me, yeah. There are times where I wonder [given our demographics…] if we were the only high school that scored in the 800s, would I wanna be recognized for that? I know it took work. You know, I know it added all these gray hairs to my head. Would I wanna have some financial recognition for that so that me and my family could take a vacation? Then I'm like, "Yeah, of course." I see it as a gravy component though; I don't see it as a driver. I see it as a recognition. I don't see it as a motivation, if that makes any sense. (School C, 232)

Dissonance

Events occurring among the leaders of the project, most often transpiring in monthly TIF steering committee meetings, almost always preceded subsequent responses among teachers. A far-reaching decision was made early on in the project that the substantial funds, paid by the federal TIF grant for capacity building around TIF metrics and implementation, were rolled over to the schools in a lump sum and were no longer available to the provider. All three schools used good portions of these funds to compensate for coincidental state budget cuts so that they could keep, or hire additional, staff or buy essential equipment. In an inducement logic, this decision made sense. As long as the recipient of funds engages in obligated practices, the funds are justified. In an incentive logic, the decision was detrimental because it made it difficult to find time to familiarize faculties with the performance management system and facilitate the

internalization of its judgments. Instead, introduction to the new system was curtailed to a few short presentations during faculty meetings, with the result that many teachers felt left in the dark about the complexity of system.

Dissonance also set in among the local TIF-leaders when it became apparent that data processing and data dashboard design were tasks much more demanding than envisioned. The technical assistance vendor seemed ill-equipped to deal with this complexity, yet no funds were available to procure additional services The result was that principals were left to carry a large part of the technical side of the performance management system, a role that absorbed all energy for TIF almost to Year 3. But the technical side of the system had to be in place to pass muster with the federal auditors. In addition, the whole performance management system consisted of 26 measures applied to the many varied roles staff play in schools. To find metrics that would create equitable opportunities for rewards across all these varied job responsibilities was a task at which the TIF steering committee eventually failed. But in the early stages, the task resembled one of fighting a many-headed serpent, and the leadership was still confident that it could prevail. By comparison, summative and formative evaluation seemed technically simple in terms of data processing, but turned out to be highly complex in coping with human judgment.

In two schools, different principals took over during the state of *dissonance* and soon came to guide the implementation of TIF. In School A, a new co-principal in charge of TIF appeared to be unwilling to support TIF, all the way to the point of refusing basic compliance with formal procedures. In School B, the new co-principal was simply overwhelmed with multiple duties. These events diminished implementation quality, and it raised the specter of financial opportunism taking over performance management intent.

Evaluation of Teaching

While in the early phases, teachers were primarily concerned about the clarity of the metrics and the process, the situation changed dramatically after the first data dashboard release in the fall of 2012 when the first summary performance scores were released and teachers had to cope with performance judgments. The scores for the formative evaluations of teaching (FET) were drawn from the quarterly conferences with instructional supervisors. Supervisors partly drew FET scores from classroom observations, but holistic judgment was their main base. As a result FET scores were more in line with teachers' expectations. Specifically, teachers commented:

> That's always my most helpful thing is when he observes me, so to get him in here on a set schedule I think was definitely helpful and I always learn from that. (School C, 220)
>
> Usually I feel it's accurate. Of course when only 20minute observations are being done, sometimes I feel maybe there was something missed or whatever, but in the most part, yeah I think. (School B, 309)
>
> I think your individual observational experiences, I think those are definitely helpful because it's like down to basics. (School A, 422)

SET scores were summative and externally generated. FET scores were on the whole, higher than SET scores. When SET scores were released, they were considered conspicuously low, despite the fact that 61 percent of participating teachers received a score of "applying" that qualified them for a monetary award. For some participating teachers, a score below 3 was not troublesome because they considered themselves novices or learners:

> The things that are highlighted as problems or things that I need to work on are the same things that I was thinking oh, yeah, that didn't go well. (School B, 303)
>
> Yeah, I did read over the feedback that I got. And that was part of actually, like, my – oh, I should focus on closing because I'm not doing one and so I did look at it. (School B, 304)
>
> I'm a brand new teacher, this is my third year. So even telling me: you're not very good at things, well, yeah, I know I'm not very good at things. (School C, 202)

But for more senior teachers who considered themselves effective or had been rated effective by FET metrics in the past, for example judgment by their instructional supervisors, the SET scores provoked disbelief:

> When people didn't get the highest set score that are used to be regarded as model teachers here or had model classrooms before, that people's SET scores seemed to be more of a discrepancy or not in line with the [school leaders]. That's where I saw more of the hurt. (School C, 220).

Two reactions to the performance pattern could have been conceivable: teachers watching anew their videos and trying to understand their score; or discarding the score as invalid and the whole process as spurious. The latter was the prevalent response.

A host of reasons were mentioned as justification for the SET scores' lack of validity and usefulness. Interviewees stated that:

* Their teaching was not aligned with the SET tool because their teaching was constructivist, a position held by many senior teachers especially at School C:

> I don't want [the SET] for science. I want it to be like inquiry style version of a lesson. Because my worry is – and actually I feel like this has become true is that whatever we push with this assessment, then becomes kind of a default practice at the school. (School C, 221)

* The lesson structure encouraged by the SET was so basic as not to deserve attention:

> I guess I'm not saying it's necessarily a useful tool […] I think the research on effective teaching and effective teaching practice is pretty clear, and these are components that should be present in your instruction. And I think that the SET is pretty aligned to that research I guess is what I'm saying. It doesn't feel like I have to do anything different for SET beyond what I would normally do when I'm teaching. (School C, 120)

* External judgment was out of context:

> I just tried to reread it and see how – in my mind, I wanted to see how – what areas did I need to improve? And then of course, I questioned it, like what are they using to measure me? Like what are the bases they're coming up with to measure me? Like it's when they're not even in my class. They're just seeing from a video. (302, School B).

* Teaching a lesson to a formal template resulted in an act of ,what Stephen Ball (2009) has called, performativity:

> I didn't – you know, my thing was like I didn't want to create like a play, like I was rehearsing for a stage thing. Because I was thinking, you know, how useful is it for me to do a fake stage that like I wasn't interested in that. (School A, 421)

* The instrument was too complex with its many behavioral indicators so that one could not follow it. There was not enough information to interpret the ratings:

> I don't find rating myself for being rated as in applying or implementing whatever the levels are of teachers to be very meaningful because I honestly don't have a strong feel for the distinctions between the different levels. So I kind of just felt like yeah, that sounds good. (219, School C).

* Without more detailed formative feedback, the summative score was useless:

> So far, I don't feel like the feedback has been necessarily that informative. I'd like to sit down and have a – like whoever observed the video to then sit down and have a conversation with me. Have a coaching session with me based on what they saw. I would take free coaching in a minute even though it means being observed and evaluated. But to get a piece of paper that says these indicators were present is not as useful to me. (School B, 301)

* The process was abstract, de-personalized, and alienating:

> And so, fine, if you want me to type it into a paper, I will, but I'm doing it for you; I'm not doing it for me. If you were going to sit down with me beforehand, talk me through it, observe me, discuss it afterwards, okay, then I have a chance to benefit from that. But if it's just me emailing you something so that you can say, "We have this many teachers that did this well in the SET lesson," like, that's not really for me. (School C, 223)

* The evaluative judgment was upsetting:

> I care about being a good teacher. And this year, my score was really low. It was lower than any – like my first year teaching, which, in my head, I was like well, this definitely isn't a valuable tool to me because I have definitely improved as an educator since my first year of teaching. So I think the tool – I get nervous and I feel vulnerable because I'm being evaluated, so it makes me feel vulnerable. (309, School B)

In sum, the prevailing sentiment was to question clarity, validity, fairness, and usefulness of the evaluations, and with it the need to learn from the information which the tool or the evaluations could potentially provide vanished. When it came time to ramp up for the Year 2 SET submissions, the skepticism towards the summative evaluation had spread across faculties and became a collective stance that expressed itself in repetitive commentary that the SET was largely invalid:

> I don't trust the SET because there's not like a connection between the person – like who is this person giving me feedback? (School B, 309)

Principals, as well, sensing their teachers' negative attitudes, did not press the point and backed off the SET by remaining silent. In a steering committee meeting, one principal, looking back, asserted that she made a point of not highlighting the system (during the phase of

dissonance) because she herself did not feel anymore that the system yielded reliable information and she did not want to upset her faculty. Participating in the SET was voluntary and the principals' silence reinforced teachers' discretion in retreating from the summative evaluation. SET video submissions declined during the dissonance phase, as shown in Table 7.

Table7: Number of Teachers Eligible for SET and Number/Percent that Submitted SET

	2011-12		**2012-13**		**2013-14**	
	Eligible for SET Reward	Submitted SET Video	Eligible for SET Reward	#Submitted SET Video	Eligible for SET Reward	#Submitted SET Video
School A	10	80% *(n=8)*	17	76% *(n=13)*	8	100% *(n=0)*
School B	10	80% *(n=8)*	10	70% *(n=7)*	14	71% *(n=10)*
School C	51	63% *(n=32)*	56	27% *(n=15)*	51	81% *(n=42)*

Those who participated became strategic:

> I put in some effort to make sure I would get the 3. So, I taught a certain kind of lesson, which I thought would ensure me that number. It wasn't hard work, it was – but it was considered work, it was strategic, "I'm going to teach this kind of lesson, because I think that kind of lesson scores this." (School B, 301)

> Oh, I just feel like it's something I have to do. So, my feelings – well, it's not really a feeling, it's like, "Oh, I better do that." Like, I'm like, "Oh, yeah, I have to remember, I have to film." (School B, 304)

> Right now it just seems like something we have to do at the end of the year. I feel like a student scrambling to get things done. (School C, 212)

Bonus Money

The way the three schools dealt with evaluations mirrored the way they dealt with bonus monies. That is, over time acceptance dwindled. While in Year 1, money had the aura of a certain innocence (who can argue with additional money?), in Year 2 money was still welcome but also an annoying and discordant feature of organizational life:

> In general everybody seemed confused. They were like okay, that's cool I got money, but I don't really understand. (School C, 220)

> I don't even know the whole matrix calculation (School B, 306)

> I think it's just like some based on like luck and other things based on like just how scores were calculated by the state. (School B, 304)

> I think people feel like whether they earn their full bonus or not is largely not dependent on what they as an individual do on a day by day basis. (School C, 217)

In the *consonance* phase, bonus monies were met with a sense of universal deservingness. If reward expectancy is the logic through which money presumably moves teachers to increase performance, this logic was completely undermined by the faulty messaging of the system. Interviewees hinted at their suspicion, or their knowledge, that payouts were surprisingly unequal across groups of teachers. Principals reported in the TIF steering committee that it was readily transparent that teachers teaching 12th grade courses were advantaged because their bonuses depended largely on internal school measures while teachers in lower grade levels were assessed on student performance on state tests. Evaluative judgment from teaching evaluations were the greatest sting:

> […] People were, like, great we got extra money, but was it worth all this extra stuff that we've all been talking about and doing? That's generally how people were acting. Some people seemed put off like I have this traditionally high status and I got one of the lowest amounts. […] What's going on with that? Even with the SET ratings I actually saw people get more upset about their SET scores than about the money itself. When people didn't get the highest SET score that are used to be regarded as model teachers here or had model classrooms before, that people's set scores seemed to be more of a discrepancy […] That's where I saw more of the hurt. The money thing was like we got this money whatever. (School A, 407)

With bonus money increasingly disconnected from performance and professional learning, expenditures of funds for bonuses *and* evaluations seemed less and less justified. Teachers, like this experienced teacher, wanted to learn, but the resources were presumably spent on things only tangentially related to learning:

> I was, like, please tell me how to do a better lesson closing. Please. I would love it. I know that my lesson closings are not effective because my class gets so differentiated that at the end I'm like: oh, God. Not once, zero times, has somebody said to me how about I watch five of your lesson closings and then you and I sit down together for an hour and talk about an alternative which you try. Then, I'll watch again. Any money that's being spent on SET. Are you kidding me? Pay that person to come do that with me. I would love that. (School C, 223).

In the dissonance phase, bonus money became a submerged topic of communication. Teachers revealed that administrators encouraged teachers to use discretion. Administrators, in the steering committee, confirmed: "We never spoke about the money. We made sure that the money issue never made it big. We knew what was behind all of this …. and so we kept quiet." (Principal School B). Other principals sheepishly concurred with this comment. As one teacher stated: *Well, they told us not to ask each other about it* (408). Teachers may have had *little side conversations* with close colleagues. But the modal answer was: *I did not talk about it with anyone.* Money, in contrast to teaching, which fell under norms of visibility, was treated as a personal issue and private in essence: *I think that the money thing has always been a private thing* (306). Some teachers were careful avoiding *comparisons,* being *sensitive of people getting different amounts,* and *being careful about making people feel bad*. Some teachers do not feel comfortable with this strategy to deal with money and see it as potential source of conflict. As on teacher explain:

> I think right now it's been kind of like this air of secrecy and privacy where like we haven't necessarily been encouraged to talk about [the scores and the money]. I think that's been a little damaging. … Don't gloat or ask people about [it] is what we say to the kids, right. Like they'll tell you if they want to tell you. It kind of created this environment, I think, where people were reluctant to talk about [it]. (217)

This feeling of discomfort was punctured when the notion spread among teachers that TIF was just one big "piñata." The notion originated in the TIF leadership or steering committee convened by the provider. Two things came together. In the transition to the common Core Standards, the state government had abandoned its state test and the NCLB-like sanctions regime. The state performance measures, however, were the linchpin of the TIF bonus pay system and were now no longer under serious consideration. The 'coup de grace' came when the bonus monies were paid on time in the fall, but because of glitches in the data base, the performance data were not released until months later, making the glaring disconnect between

bonus award and performance visible to all. The notion of a "piñata" took the sting out of the differential amounts that the TIF system bestowed on teachers. It returned the system to its original 'innocence' during the period of consonance when it was unconnected to performance and was embraced as bringing money to the school in whatever form and as validating teachers' deservingness of reward.

Resonance

Resonance sprouted when teachers in the three schools had found ways to insulate themselves from evaluative discomfort and the divisionary effects of differential pay for unjustified performance. As in the two previous periods, the new development originated in the TIF leadership team. The "piñata" moment had brought provider, school administrators, and evaluators together in jointly acknowledging that the incentive function of the performance management system had been a failure. The state tests had vanished and the NCLB sanctions regime had imploded. The technical side of the TIF system was in shambles. Participation in the SET, the only part of the system that had any viability, had shrunk. Concurrently a new director took the helm of the provider organization. At its nadir, the leadership changed course. From now on, a portion of every steering committee meeting was dedicated to analyzing videos, submitted by teachers for summative evaluation.

Evaluation of Teaching and Bonus Money

In year three of implementation, the incentive and performance management function of the TIF system was thoroughly discredited, and it was overtly rejected. For the money, the piñata theme was repeatedly invoked, and it inured teachers to potentially divisive or discomforting sentiments. Most teachers stopped interacting with the data dashboard altogether and did not check their performance scores anymore, such as these teachers:

> I think at some point you just decide that it's not valid information or something because the information that you get back on it isn't just – I don't find it to be especially useful as far as figuring out what I need to work on. (School C, 217)

> I actually find it to be punitive sometimes rather than encouraging. I think it just picks at things that are happening. (School B, 301)

Performance scores and bonus payments were treated with silence and became a largely private affair. But below outward silence, they exerted their latent presence, not as valid measure of one's performance, but as an irksome feature that could potentially disquiet one's self-perception as a good performer and one's sense of being fairly compensated for one's work relative to others. Those who had negative experiences with their scores tried to distance themselves:

> I told [my supervisor] what would work for me is if for her to see the data and then because she knows how to work with me, she knows how to be gentle or whatever, so if she reads the data and then she could help me through that data like oh, these are the things you did great. (308, School B)

Despite this distancing, the evaluation remained a latent concern:

> I'm curious. I'm curious as to you know, what they're going to say and – but I don't have like you know, big expectations. (School A, 419)

> I think with the SET last year it was like – a little bit like, oh – you know, I hadn't really necessarily tried a ton. But, you know, you'd like to, you know be rated well, and I wasn't. (School C, 221)

Participation in the summative evaluation actually bounced back from the low of Year Two as indicated by submitted videos, but this cannot be attributed to higher acceptance of the system. Rather, administrators, especially in School C, fearful of undermining the viability of TIF (and with it the flow of incentive monies) made it clear to staff that they expected participation, and teachers became more strategic in making sure that they obtained bonus awards. More than half of the teachers interviewed during the Year Three who submitted SET video submissions acknowledged that money was one reason to submit the video. Few teachers, however, admitted that they made a specific investment in the preparation of the video as a way to increase the bonus.

Teacher Learning and Artifacts

While the use of artifacts associated with teacher evaluation—namely the SET observation instrument and the videos—receded for summative and incentive purposes, their use in learning events advanced. This was largely due to the role of the provider during the 2013-14 school year. The provider invited principals and other instructional leaders from the schools to systematically analyze samples of teaching from the SET video submissions and to recognize strengths and weaknesses in submitted lessons. After viewing lessons submitted by their teachers, leaders found renewed value and interest in using the artifacts provided by TIF to improve instruction. Leaders at Schools A and B were especially receptive.

School A

At School A, the school principal and an instructional coach started a professional development series with teachers where they would use sections of the SET tool to guide each session, and would supplement these session with academic articles to deepen their understanding of the SET and emphasize the importance of student engagement and learning. As part of this work, teachers were then asked to design and lead a lesson that other teachers would observe. Through this initial professional development series, teachers then began a lesson study routine using an observation tool that was internally developed to capture general reflections from classroom visits. This tool was not related to the SET; rather, it was an open-ended observation instrument that was shaped by questions that would help the observer provide feedback on student learning and teacher instruction. After the lesson study, teachers would reconvene in the professional development sessions to debrief on the classroom observations. The last two sessions in this series were reserved for SET preparation, and teachers would collaborate on preparing the lesson plan for their final submission under the supervision of the

principal and instructional coach. As noted by the principal, institutionalizing these routines and standardizing teaching with the five-part lesson was part of the school strategy during the last phase of the TIF grant:

> I mean, we did things whole group. We did things in department teams. We reinforced it [the five-part lesson plan]. Those of us who do observe and coach teachers in different capacities reinforced it in those coaching conversations. And I really learned a lot this year from the provider on how to have those conversations that are more focused on, okay, here's the model. Where do you see alignment or not alignment between what you did in the model? (School A, 423)

Teachers at School A also began using the five-part lesson plan to shape their lessons throughout the year, and they expressed value in using the revised SET tool as part of their lesson study professional development:

> The five-part tool? Yeah, to an extent, right. I think really understanding – I think I really grasped that five-part lesson plan last year, and so it was something that was very familiar with me. And something that I've been trying to do all year really is really instill that five-part lesson plan explicitly in every lesson. Like really focusing on modeling, really making sure that I'm checking for understanding, really focusing on formative assessments, really focusing on engagement. (School A, 416)

> I do use the SET information…It was much more the act of filming yourself and specifically making sure this is where your hook is. Do you have a closure piece in there? Like do you have some check for understanding? Did you make an exit ticket? Do you have some sort of formative assessment? Like deliberately making sure that all those parts are in line was the thing that I think improved my practice the most. More than like the list of things you should have. Does that make sense? (School A, 425)

School B

In School B, attention to the clinical nature of the SET had already begun in the spring of 2013 ahead of the other two schools. Led and organized by the school's instructional coach, the practice of lesson study was structured focusing on different components of the SET tool (introductory phase, modeling, closing phase, etc.). Each teacher would film the given segment of their lesson by themselves, and then teachers would share the lesson segment they filmed during inquiry group meetings. Teachers were encouraged to share their lesson plan in advance of the meeting to have some context for the videos, and then teachers would provide feedback to

one another after viewing the footage together using the SET instrument to frame their feedback. The instructional coach at School B encouraged this routine early on as a way to give teachers space for collaboration and feedback:

> Well, it did start from having to do the SET, right, a few years ago. So thinking of if the lens were looking at teaching and right now it's like the five part lesson, giving space or an opportunity for teachers to really closely analyze their own work and each other's work and have a chance to collaborate in terms of how to make our practice more effective so that students are learning. I mean, to just hone in on the teaching and learning. It could be done through the five part lesson. It could be done through other things, but actually looking at practice. Having space for teachers to talk about practice. (School B, 301)

Teachers also showed appreciation for the lesson study process, and how it would help in their teaching and final SET submission:

> But like last year, we're doing the lesson study again this year. So we're recording ourselves, different parts of the lesson, and watching it in our little small groups and giving each other feedback. That is useful. So the SET in a way is a culmination of that…In a way, maybe the SET will feel more like a culmination. Before, it was like start. Pause. Time, time goes by. Do the SET. Now, it will be like we've really looked at opening and modeling phases. Then next month, we're going to really look at like building and checking for understanding. Good. That's the one I really want to look at with my colleagues. Then, hopefully by the time we do the SET, it will feel more integrated. (School B, 301)

School C

The story of School C is different from the other schools. School C had started using the SET during their inquiry groups as early as 2012, but the instructional coaches felt strongly that the SET tool would need to be modified to better suite the needs of their teachers, many who taught with a constructivist lesson framework rather than one of direct instruction. The coaches pursued a revision of the SET in fall 2013 to include the option for a constructivist lesson plan. Working with the provider and the program evaluators, a new option for a constructivist lesson was inserted into the SET instrument. These changes led to more acceptance of the artifact at School C, and the instructional leaders rolled-out the revised instrument in the spring of 2014 at a school-wide professional development meeting. By being responsive to the schools' needs, the revised instrumented helped leaders imbed the tool into their organizational culture in a deeper

way by continuing to use it in inquiry groups, and also began using it as a way to train new teachers:

> I definitely used the revised tool this year in a couple of forms. One is most basic starting point, I used it as a baseline for inquiry group to think about just the lesson components that need to be there in terms of kind of backed into it in terms of talking about what needed to be there for their SET, but it allowed for my repeated reminder that it's good instruction. That we need an opening and a middle and we need to give feedback and check for understanding in the middle and have a closing, and that's what we need to do, so I use it in that way. (School C, 221)

> I ran one coaching cycle with a brand new teacher where we looked at – actually used the tool and the indicators to evaluate someone else's openings for two weeks of lesson planning and that was pivotal for her. Like that experience of like of going through and saying oh, this opening engages students. This opening, wow, it engages students and it was retrieval. Like that, just that process was really beneficial for her to see to use it in terms of seeing someone else's practice. (201, School C).

Teachers also showed appreciation for the revised SET since it was better aligned with some teachers' constructivist pedagogical style and discipline:

> When [the instructional coach] was talking to us, he was using SET kind of like a guideline to talk about our lesson study. As a science department we did a lesson study, but we used kind of like the guidelines were SET like this is an opening, this is like a closing. Like we used those things in planning the lesson study lesson, and so for maybe three weeks, four weeks, three to four weeks like in our inquiry group meetings we kept going back to that to use the plan what the lesson study was going to look like. And then we all went into a chemistry class and [the instructional coach] led the lesson and the department teachers were in there just getting data. But we used the SET to like plan for that lesson because we used it as preview to get us ready, we had to do it ourselves. (School C, 241)

> Here I actually used a lesson like you normally do in science during the SET, because before it had to be a five-part lesson plan, and we don't do a lot of direct instructions in science right. (School C, 202)

Despite this renewed energy at School C, some teachers were still concerned that the tool was not going to provide teachers with the feedback specificity they were hoping for, nor was it deeply institutionalized into the school norms:

> I'm not against the idea of the SET. I know it really probably sounds like I am. I think if we could figure out a way to make it something that provides meaningful feedback and something that then is really used by coaches to shape goals or guide future practices – but that requires so many things. It requires that the feedback received as something that coaches actually think is valuable and educators actually think is valuable. And it has to, as you observed, come back in a reasonable amount of time. (School C, 217).

Discussion

The purpose of the paper is to understand the ways components of the TIF architecture, namely teacher evaluations and bonus payments, mandated or regulated by external agencies (e.g., the federal TIF program and the local TIF provider) influence teacher learning around instruction. We pursue this purpose by untangling dynamics related to evaluation, rewards, and practices obligated by tools or artifacts. As we trace their interplay, we tease out the unique contributions of each of these elements to the effect of the system as a whole. We saw from the literature that for proponents of incentive-driven performance management of the TIF type, the various elements fuse: powerful evaluation tools spell out what matters for good teaching and diagnose the gap between desired and actual performance. Evaluative judgment compels teachers to take the need to learn and improve seriously, and rewards additionally spur teachers to strive towards desired outcomes. But do those elements, indeed, fuse or can we identify distinct dynamics?

In a nutshell, we saw that the interaction of evaluation, bonus pay, artifacts obligated by the system, and teacher learning around instruction shifted over three distinct periods that we called *consonance, dissonance,* and *resonance.* In the consonance phase, all elements fused, but this fusion was dependent on a selective perception and interpretation of what the system was all about. In the dissonance phase, the elements become discordant to each other and to the established culture of the schools. Focus on learning became crowded out as teachers' and administrators' concerns about the system's incentive function took over. In the resonance phase, the pattern shifted to the opposite. The incentive function (summative evaluations, bonus pay) were largely rebuffed and relegated to the periphery. Once the power of incentives became latent, first administrators and then teachers came to interact with the two main artifacts, videos

of lessons and the clinically oriented SET observation tool, in novel ways. In conjunction with professional development supplied by the provider for the instructional leaders, the videos and the clinical observation tool became resonant with internal concerns for high quality teaching. The system-induced artifacts reconnected to the initial concern for precise feedback. Once the tools were overtly decoupled from incentives, they could become learning tools, though the obligation to engage with them to secure federal money, and the evaluative discomfort caused by external evaluation and differential bonus awards remained latent forces.

Yet, resonance does not mean that the tools and practices obligated by TIF became powerful drivers of teacher learning. They pricked the surface and began to seep into established practices. In two schools, as a result of the instructional leaders' initiative, the tools advanced the *idea* of focusing attention on lessons and on studying the evidence from lesson delivery. One school adjusted the observation tool and downshifted the clinical precision of their learning. Another school translated the *idea* of clinical observation into an open, non-clinical inquiry format that had been their traditional approach. In a third school, only some instructional coaches picked up the clinical tool for the training of their beginning teachers. Without the signal of the administration that the school as a whole would invest in the approach and the SET tool, the practices remained confined to pockets. It is important, however, to note that even resonance of a somewhat tenuous nature was impossible until faculties in the schools had blunted the incentive function of the performance management system. Evaluative judgment and differential reward stressed the constraints imposed on teachers by the TIF-obligated practices. It was not until summative evaluation and differential reward were attenuated that the tools' affordances for learning could come into view and reconnect to initial desires to learn and grow.

We see two sub-strata in the data, one related to formative judgment and learning, the other to summative evaluation and incentives. As to learning, TIF began with the hope and articulated need among faculty that the evaluative side of TIF may contribute to teacher learning, most notably by making feedback on instruction more frequent and more precise. In the beginning, TIF was not associated with summative evaluation but with formative judgment promoting learning. This made sense given the context of an adult learning culture characterized by high visibility, collegiality, and engagement in personal growth. As to incentives, TIF began with the notion of collective deservingness of rewards. TIF monies were not interpreted as performance contingent incentives, but as inducements for good work all around. This made sense, given the strong service ethic and communal orientation of the faculty, including administrators and teachers.

Thus, the TIF incentive function had to find ways to enter educators' initial assumptions and established routines. Summative evaluation (e.g., SET scores) and differential incentives (bonus awards) needed to anchor in the sub-stratum of collective deservingness and formative judgment. The literature on performance management, evaluation, and incentives, reviewed earlier, would predict some dissonance in this process. But under the right kinds of circumstances, summative evaluation and formative judgment may fuse, and differential bonuses may become markers of competence and deserved reward. If this happens, their separate functions are hard to disentangle. Not so in our case.

Our three cases afford us a view of a performance management system in which the various functions played out in a strongly distinct fashion over time. Bonus money was welcome any time, but never rose to a distinct reward calculus that connected to meaningful tasks, such as increasing teaching excellence. Summative evaluation decoupled from formative feedback and

learning and with it the use of tools decoupled from teacher learning as well. Desire to learn and grow as instructors in more precise and clinical ways was crowded out intermittently. If anything, bonuses placed on evaluation scores diminished the chances of learning from evaluation. Learning could re-attach itself to the clinical tools once the effects of incentives were buffered. Rather than interacting synergistically, the various elements of the system seemed to be in conflict with each other.

Why was this the case? The theoretical literature offers several salient explanations:

- Poor implementation quality: our data show that data management was complex and the construction of fair metrics eluded the local TIF leaders. Moreover, capacity building around the new metrics was very limited.
- Reward expectancy: our data show that teachers found it hard to connect monetary rewards to effort and performance.
- Clarity and procedural or distributive fairness: our data show that summative evaluations were viewed as low in all these respects.

These are explanations that are clearly applicable to our data. But explanations have to go more deeply. For low implementation quality was only partly due to technical difficulties, and especially teaching evaluations did not suffer to the same degree from these difficulties. Poor implementation is attributable to decisions taken by leaders who from the start saw TIF as a way to garner resources and were averse to using it as a performance management system. The teachers articulated two separate unconnected needs or desires: to reap additional money and to receive feedback for learning in more precise ways. When clinical observation tools that could have facilitated this learning became tainted and associated with incentives, the tools were rejected.

Throughout all phases, whether consonance, dissonance, or resonance, the clinical tools of the system, the SET, the videos, and the formative quarterly conferences, were always there as obligations to be engaged with, and as latent nuisance to justify the flow of money. Over time, they seeped into faculties' shared cognitions. Befitting the established culture of the schools, TIF induced behaviors, it never incentivized them. Once the TIF obligation is gone and the money no longer available, would the use of the system's tools continue and the idea of clinical precision maintain its hold?

Bibliography

Archer, J., Kerr, K., & Pianta, R. (2014). Why Measure Effective Teaching. In T. Kane, K. Kerr, & R. Pianta (Eds.), *Designing Teacher Evaluation Systems: New Guidance from the Measures of Effective Teaching Project*. San Francisco, Ca: John Wiley and Sons, Inc.

Aucoin, P. (1990). Administrative reform in public management: paradigms, principles, paradoxes and pendulums. *Governance*, *3*(2), 115–137.

Balse, J., & Kirby, P. (2008). *Bringing out the Best in Teachers: What Effective Principals Do*. Thousand Oaks, Ca: Corwin Press.

Barzelay, M. (2001). *The New Public Management: Improving research and policy dialogue*. Berkeley: University of California.

Beer, M., Cannon, M. D., Baron, J. N., Dailey, P. R., Gerhart, B., Heneman, H. G., … Locke, E. A. (2004). Promise and peril in implementing pay-for-performance. *Human Resource Management*, *43*(1), 3–48.

Bill and Melinda Gates Foundation. (2013). *Ensuring Fair and Reliable Measures of Effective Teaching*. Seattle, Wa: Bill and Melinda Gates Foundation.

Blase, J., & Blase, J. (1998). *Handbook of Instructional Leadership: How Really Good Principal Promote Teaching and Learning*. Thousand Oaks, Ca: Corwin Press.

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

Cole, M., & Engstrom, Y. (1993). A cultural-historical approach to distributed cognition. In G. Salomon (Ed.), *Distributed cognitions* (pp. 88–110). New York: Cambridge University Press.

Danielson, C. (2011). *Enhancing Proffessional Practice: A Framework for Teaching* (2nd ed.). Alexandria, Va: Association for Supervision and Curriculum Development.

Darling-Hammond, L. (2013). *Getting Teacher Evaluation Right*. New York, NY: Teachers College Press.

Darling-Hammond, L., Wise, A., & Pease, S. (1983). Teacher Evaluation in the Organizational Context: A Review of the Literature. *Review of Educational Research*, 285–328.

Debray, E. H. (2006). *Politics, Ideology & Education: Federal Policy During the Clinton and Bush Administrations*. New York: Teachers College Press.

Donaldson, S., Gooler, L., & Scriven, M. (2002). Strategies for Managing Evaluation Anxiety: Toward a Psychology of Program Evaluation. *American Journal of Evaluation*, 261–273.

Ferlie, E., Ashburner, L., Fitzgerald, L., & Pettigrew, A. M. (1996). *The new public management in action*. Oxford university press Oxford.

Finnigan, K. S., & Gross, B. (2007). Do Accountability Policy Sanctions Influence Teacher Motivation? Lessons From Chicago's Low-Performing Schools. *American Educational Research Journal*, *44*(3), 594–630. http://doi.org/10.3102/0002831207306767

Frey, B. S., Homberg, F., & Osterloh, M. (2013). Organizational Control Systems and Pay-for-Performance in the Public Service. *Organization Studies*, *34*(7), 949–972.

Gery, G. (1991). *Electronic performance support systems*. Tolland, MA: Cery Associates.

Glazerman, S., McKie, A., Carey, N., & Harris, D. (2012). *Evaluation of the Teacher Advancement Program (TAP) in the Chicago Public Schools: Study Design Report*. Chicago, IL: Joyce Foundation.

Glickman, C. (2002). *Leadership for Learning: How to Help Teachers Succeed*. Alexandria, Va: Association for Supervision and Curriculum Development.

Goldstein, J. (2007). Easy to Dance to: Solving the Problems of Teacher Evaluation with Peer Assistance and Review. *American Journal of Education*, 479–508.

Gregoire, M. (2003). Is it a challenge or a threat? A dual-process model of teachers' cognition and appraisal processes during conceptual change. *Educational Psychology Review*, *15*(2), 147–179.

Hallinger, P., Heck, R. H., & Murphy, J. (2014). Teacher evaluation and school improvement: An analysis of the evidence. *Educational Assessment, Evaluation and Accountability*, *26*(1), 5–28.

Halverson, R. (2003). Systems of practice: How leaders use artifacts to create professional community in schools. *Education Policy Analysis Archives*, *11*(37), 1–34.

Halverson, R., & Clifford, M. (2006). Evaluation in the wild: A distributed cognition perspective on teacher assessment. *Educational Administration Quarterly*, *42*(4), 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.

Harackiewicz, J. M., & Sansone, C. (2000). *Rewarding competence: The importance of goals in the study of intrinsic motivation.* Academic Press.

Hatry, H., Greiner, J., & Ashford, B. (Eds.). (1994). *Issues and Case Studies in Teacher Incentive Plans*. Washington, DC: The Urban Institute Press.

Hill, H., Kapitula, L., & Umlad, K. (2010). A Validity Argument Approach to Evaluating Teacher Value-Added Scores. *American Educational Research Journal*, 794–831.

Hood, C. (1991). A public management for all seasons? *Public Administration*, *69*(1), 3–19.

Hutchins, E. (1995). *Cognition in the Wild*. Cambridge, MA: MIT Press.

Hutchins, E. (2006). The distributed cognition perspective on human interaction. In N. Enfield & S. Levinson (Eds.), *Roots of human sociality: Culture, cognition and interaction* (pp. 375–398). New York: Berg.

Ingersoll, R. M. (2001). Teacher turnover and teacher shortages: An organizational analysis. *American Educational Research Journal*, *38*(3), 499–534.

John, P., & Sutherland, R. (2005). Affordance, opportunity and the pedagogical implications of ICT. *Educational Review*, *57*(4), 405–413.

Johnson, S. M. (2007). *Finders and keepers: Helping new teachers survive and thrive in our schools.* ERIC.

Kane, T., Kerr, K., & Pianta, R. (2014). *Designing Teacher Evaluation Systems: New Guidance from the Measures of Effective Teaching Project*. San Francisco, Ca: John Wiley and Sons, Inc.

Kimball, S. (2002). Analysis of Feedback, Enabling Conditions and Fairness Perceptions of Teachers in Three School Districts with New Standards-Based Evaluation Systems. *Journal of Personnel Evaluation in Education*, 241–268.

Kluger, A., & DiNisi, A. (2006). The effects of feedback intervention on performance: Historical review, meta-analysis, a preliminary feedback intervention theory. *Psychological Bulletin*, 254–284.

Knapp, M., & Feldman, S. (2012). Managing the intersection of internal and extrnal accountability: Challenge for urban school leadership in the United States. *Journal of Personnel Evaluation in Education*, 664–694.

Lazear, E. P. (2003). Teacher incentives. *Swedish Economic Policy Review*, *10*(2), 179–214.

Louis, K., Marks, H., & Kruse, S. (1996). Teachers' Professional Community in Restructuring Schools. *American Educational Research Journal*, 757–798.

Marsh, J., Springer, M., McCaffrey, F., Yuan, K., Epstein, S., Koppich, J., … Peng, A. (2011). *A Big Apple for Educators New York City's Experiment with Schoolwide Performance Bonuses*. DTIC Document.

Max, K., Constantine, J., Wellington, A., Halgren, K., Glazeman, S., Chiang, S., & Speroni, C. (2014). *Evaluation of the Teacher Incentive Fund: Implementation and Early Impacts of Pay-for-Performance After One Year*. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, US Department of Education.

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

Milanowski, A., & Heneman, H. (2011). Assessment of Teacher Reactions to a Standards-Based Teacher Evaluation System: A Pilot Study. *Journal of Personnel Evaluation in Education*, 193–212.

Milanowski, A., Heneman, H., & Kimball, S. (2011). *Teaching assessment for teacher human capital management: Learning from the current state of the art*. Working Paper.

Milanowski, A. T., & Heneman III, H. G. (2001). Assessment of Teacher Reactions to a Standards-Based Teacher Evaluation System: A Pilot Study*. *Journal of Personnel Evaluation in Education*, *15*(3), 193–212.

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

Millman, J., & Darling-Hammond, L. (1990). *The New Handbook of Teacher Evaluation: Assessing Elementary and Secondary School Teachers*. Thousand Oaks, Ca: Corwin Press.

Mintrop, H. (2004). *Schools on probation: How accountability works (and doesn't work)*. Teachers College Pr.

Mintrop, H., & Sunderman, G. L. (2009). Predictable failure of federal sanctions-driven accountability for school improvement—and why we may retain it anyway. *Educational Researcher*, *38*(5), 353–364.

Mitchell, D. E., Ortiz, F. I., & Mitchell, T. K. (1987). *Work orientation and job performance: The cultural basis of teaching rewards and incentives*. SUNY Press.

Murnane, R. J., & Cohen, D. K. (1985). Merit Pay and the Evaluation Problem: Understanding Why Most Merit Pay Plans Fail and a Few Survive.

Murphy, J., Hallinger, P., & Heck, R. (2013). Leading via Teacher Evaluation: The Case of the Missing Clothes? *Educational Researcher*, 349–353.

National Research Council. (2011). *Incentives and Test-Based Accountability in Public Education.* Washington, DC: National Research Council.

Nemeth, C., Cook, R., O'Connor, M., & Klock, A. (2004). Using cognitive artifacts to understand distributed cognition. *IEEE Transactions on Systems, Man, and Cybernetics*, *34*(6), 726–735.

Nemeth, C., O'Connor, M., Klock, P., & Cook, R. (2006). Discovering healthcare cognition: The use of cognitive artifacts to reveal cognitive work. *Organization Studies*, *27*(7), 1011–1035.

Odden, A., Kelley, C., Heneman, H., & Milanowski, A. (2001). *Enhancing Teacher Quality through Knowledge and Skills Based Pay*. Philidelphia, Pa: Consortium for Policy Research in Education.

Osborne, D., & Gaebler, T. (1992). Reinventing government: How the entrepreneurial spirit is transforming government. *Reading Mass. Adison Wesley Public Comp*.

Pianta, R., Paro, K., & Hamre, B. (2005). *Classroom Assessment Scoring System (CLASS)* (Unpublished measure). Charlottesville, Va: University of Virginia.

Podgursky, M. (2006). Teams versus Bureaucracies: Personnel Policy, Wage-Setting, and Teacher Quality in Traditional Public, Charter, and Private Schools. *Education Working Paper Archive*.

Podgursky, M. J., & Springer, M. G. (2007). Teacher performance pay: A review. *Journal of Policy Analysis and Management*, *26*(4), 909–950.

Podgursky, M., & Springer, M. G. (2006). Teacher performance pay: A review. *National Center on Performance Incentives*, 2006–01.

Pollitt, C., & Bouckaert, G. (2011). *Public management reform: A comparative analysis-new public management, governance, and the Neo-Weberian state*. Oxford University Press.

Raudenbush, S. (2004). What are Value-Added Models Estimating and What Does this Imply for Statistical Practice? *Journal of Educational and Behavioral Statistics*, 121–129.

Sansone, C., & Harackiewicz, J. M. (2000). *Intrinsic and extrinsic motivation: The search for optimal motivation and performance*. Academic Press.

Scriven, M. (1974). The Evaluation of Teachers and Teaching. *California Journal of Educational Research*, 109–115.

Shipps, D. (2006). *School reform, corporate style: Chicago, 1880-2000*. Univ Pr of Kansas.

Solomon, G. (1993). *Distributed cognitions: Psychological and educational considerations* (Vol. 11). Cambridge: Cambridge University Press.

Springer, M. G. (2009). Rethinking teacher compensation policies: Why now, why again? *Performance Incentives: Their Growing Impact on American K-12 Education, Brookings Institution Press, Washington DC*, 1–22.

Springer, M. G., Ballou, D., Hamilton, L., Le, V. N., Lockwood, J. R., McCaffrey, D. F., … Stecher, B. M. (2011). Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching (POINT). *Society for Research on Educational Effectiveness*.

Springer, M. G., Pane, J. F., Le, V.-N., McCaffrey, D. F., Burns, S. F., Hamilton, L. S., & Stecher, B. (2012). Team Pay for Performance: Experimental Evidence From the Round Rock Pilot Project on Team Incentives. *Educational Evaluation and Policy Analysis*.

Stiggins, R., & Duke, D. (1988). *The Case for Commitment to Teacher Growth: Research on Teacher Evaluation*. Albany, NY: State University of New York.

Stodolsky, S. (1990). Classroom Observation. In J. Millman & L. Darling-Hammond (Eds.), *The New Handbook of Teacher Evaluation: Assessing Elementary and Secondary School Teachers* (pp. 175–190). Newbury Park, Ca: Sage.

Supovitz, J. (2002). Developing Communities of Instructional Practice. *Teachers College Record*, 1591–1626.

Taut, S., & Sun, Y. (2014). The Development and Implementation of a National, Standards-based, Multi-method Teacher Performance Assessment System in Chile. *Education Policy Analysis Archives*, *22*(0), 71.

The New Teacher Project. (2009). *The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness*. Brooklyn, NY: The New Teacher Project.

Timperley, H. S., & Robinson, V. M. (1998). Collegiality in schools: Its nature and implications for problem solving. *Educational Administration Quarterly*, *34*(1 suppl), 608–629.

Toch, T., & Rothman, R. (2008). *Rush to Judgment: Teacher Evaluation in Public Education*. Washington DC: Education Sector.

Trujillo, T. M. (2013). The Disproportionate Erosion of Local Control: Urban School Boards, High-Stakes Accountability, and Democracy. *Educational Policy*, *27*(2), 334–359.

Tucker, P. (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*, 103–126.

Vroom, V. (1964). *Expectancy theory*. John Wiley and Sons.

Yuan, K., Le, V.-N., McCaffrey, D. F., Marsh, J. A., Hamilton, L. S., Stecher, B. M., & Springer, M. G. (2012). Incentive Pay Programs Do Not Affect Teacher Motivation or Reported Practices: Results From Three Randomized Studies. *Educational Evaluation and Policy Analysis*, *35*(1), 3–22.