IMPLEMENTING TEACHER EVALUATION SYSTEMS: HOW PRINCIPALS MAKE SENSE OF COMPLEX ARTIFACTS TO SHAPE LOCAL INSTRUCTIONAL PRACTICE[1]

Richard Halverson, Carolyn Kelley, and Steven Kimball

Abstract

This study examines how local school leaders make sense of complex programs designed to evaluate teachers and teaching. New standards-based teacher evaluation policies promise to provide school leaders and teachers with a common framework that can serve as a basis for improving teaching and learning in schools (Danielson & McGreal, 2000; Odden & Kelley, 2002). However, implementation research suggests that the ways in which local actors make sense of and use such policies determines the nature of the changes that actually occur in schools (Desimone, 2002; Spillane, Reiser, & Reimer, 2002). In this paper, case studies of schools in a large school district are used to examine school-level implementation of a standards-based teacher evaluation system. Specifically, we examine the ways in which school and district leaders emphasize and select from the many features of a teacher evaluation framework in the implementation process. We then discuss the ways in which key features of the process were co-opted, ignored, or adapted in accordance with school context, and we point to how the resulting teacher evaluation practices help to create conditions for more substantive conversations about reforming teaching practice.

Introduction

While it is generally acknowledged that teachers exert great influence over the improvement of student learning (Darling-Hammond & Ball, 1997; Wright, Horn & Sanders, 1997), the role that school leaders play in shaping system capacity for successful teaching

This paper will be published in the January 2004 edition of *Theory and Research in Educational Administration*
Please do not distribute without the first author's permission (halverson@education.wisc.edu)

and learning is often underappreciated (Murphy, 1994; Hallinger & Heck 1996; Elmore 2002). For the most part, principals affect instruction indirectly, through practices such as the acquisition and allocation of resources, supporting and encouraging staff, enforcing rules for student conduct, or taking personal interest in the professional development process (Berends, et. al., 2002; Peterson, 1989). However, principals can also affect teaching practice directly through teacher supervision and evaluation. Evaluation is a formal means for school leaders to communicate organizational goals, conceptions of teaching, standards, and values to teachers (Wise, Darling-Hammond, McLaughlin & Bernstein, 1984).

Teacher Evaluation Frameworks

Teacher evaluation is a common, often mandatory practice in schools. The traditional programs and practices of teacher evaluation, however, are based on limited or competing conceptions of teaching (Darling-Hammond, Wise, & Klein, 1999), and are often characterized by inaccuracy, lack of support (Peterson, 1995) and insufficient training (Loup, Garland, Ellett, & Rugutt, 1996). Traditional teacher evaluation practices tend to preserve the loose coupling between administration and instructional practices, consequently limiting the ability of principals to foster improvements in teaching and learning (Weick, 1976; 1996; Rowan 1990). Rather than being used as tools for instructional leadership, traditional evaluation programs are often seen as perfunctory and treated by both teachers and principals as an administrative burden. Teacher assessment has frequently been used to weed out the poorest performing teachers rather than to hold all teachers accountable or to improve the performance of all teachers (Darling-Hammond et al., 1999; Haney, Madaus & Kreitzer, 1987). Because of these traditional limits on scope and efficacy, teacher evaluation has had a

limited impact on teacher performance and learning (Peterson, 1995; Darling-Hammond, Wise & Pease, 1983).

A number of districts developed evaluation systems based on teaching standards to address these concerns. These new systems focus evaluation on a common vision of teaching elaborated across broad domains of practice, comprehensive standards and rubrics, and multiple-sources of evidence (Kimball, 2003; Milanowski & Heneman, 2001; Davis, Pool, & Mits-Cash, 2000). One such model, Danielson's (1996) *Enhancing Professional Practice: A Framework for Teaching*, develops standards to assess and promote teacher development across career stages, school levels, subject matter fields, and performance levels. The framework is organized into four domains of planning and preparation, the classroom environment, instruction, and professional responsibilities. These domains include 22 components spelled out by 66 elements to specify a range of appropriate behaviors. Each element includes rubrics to assess unsatisfactory, basic, proficient, and distinguished performance. The framework is also intended to foster teachers' development by specifying techniques for assessing each aspect of practice, a program of evaluator training, and emphasis on using the framework to include formative as well as summative evaluation (Danielson & McGreal, 2000).

Prior research on the implementation of this type of standards-based teacher evaluation system has examined the initial perceptions of teacher and administrator acceptance (Milanowski & Heneman, 2001; Davis, Pool, & Mits-Cash, 2000), the nature of feedback, enabling conditions and fairness perceptions (Kimball, 2003) and the relationship of these evaluation systems to student achievement (Gallagher, 2002). Yet we know relatively little about how local school leaders actually use such systems in practice, which

features they select from the frameworks to emphasize in their evaluations, and how they adapt the systems to existing evaluation practices. In this paper, we use sensemaking theory as a lens to examine how local school leaders use the framework to shape teaching practices in schools. This knowledge will help policy makers and school leaders to better understand both obstacles and opportunities afforded by comprehensive teacher evaluation frameworks.

Sensemaking

Sensemaking theory addresses the cognitive dimensions of change in people and organizations (Spillane, Reiser, & Reimer, 2002; Weick, 1996). Sensemaking begins with the constructivist assumption that learning is shaped by prior experience (Greeno, Collins, & Resnick, 1996**;** Confrey, 1990). Through experience, people build mental models to anticipate regular patterns of action in the world (Gentner & Stevens 1983; Hammer & Elby 2002). Mental models act as perceptual filters that help to determine both what we notice, and how it is interpreted (Starbuck & Milliken, 1988). Our models shape what we notice in new experiences, and can override the potential of new ideas to transform behavior (Cohen & Barnes, 1993). The impact of new ideas can be either marginalized or co-opted by preexisting practices and ideas (Chinn & Brewer, 1993; Keisler & Sproull, 1982). The tendency to interpret the new in terms of the old may lead people to attend to the surface similarities of new concepts and practices instead of attending to the deeper, structural differences (Gentner, Ratterman & Forbus, 1993; Ross, 1987). People also tend to retain practices they value, and value the practices they retain.

The sense we make of new information is also shaped by our social and situational context (Greeno, 1998). Organizations and institutions routinize existing models through policies, programs, and traditions. Thus, the intended effects of innovations are not

necessarily altered by the malice or laziness of implementers, but instead by the best efforts of local actors seeking to satisfice conflicting goals (Spillane, Reiser, & Reimer 2002, Fischoff 1975; March & Simon, 1958). Actors make sense of new practices within their existing social and situational context, and often adjust the meaning of the new in terms of their established context of meaning.

Our cognitive models, however, are not rigid structures that determine what we notice and name. Rather, our models interact with our perceptions and experience in an iterative process through which new experiences can come to shape our existing models. Successful learning requires an active process of readjusting mental schema to what we already know (Carey, 1985; Schank & Abelson, 1977). The tenacious hold our existing ideas have on what we notice and name can require an experience of expectation failure to jolt us into reconstructing our network of assumptions (Schank, 1982). In organizations, new policies and programs can provide this jolt to existing practice, encouraging practitioners to reframe their practice in terms of the new expectations. The ways that practitioners make sense new initiatives in terms of pre-existing models make the implementation of new, complex programs a far from linear and predictable process.

Artifacts as a Window on Sensemaking

Because of its iterative and transitory nature, the sensemaking process has proven to be difficult to research. One way to access sensemaking is to identify occasions when existing models are perturbed by interventions (Bronfenbrenner, 1979). Leaders and policy-makers introduce policies and programs into organizations to reshape existing practices. In these cases, policies and programs can be understood as sophisticated *artifacts* intended to shape or reform existing practices in an institutional context (Pea, 1993; Norman, 1993;

Wartofsky, 1979; Halverson & Zoltners, 2001). Organizational artifacts originate from different locations. Artifacts such as district policies, state and federal programs, and teacher professional networks originate outside the local school context, whereas other artifacts originate within the school as locally designed efforts to resolve emergent and/or recurrent problems of practice (Halverson, 2002). Taken together, the network of received and locally designed artifacts composes a local situation that both facilitates and constitutes local leadership and teaching practice (Spillane, Halverson & Diamond, 2001).

Artifacts have several features important for understanding sensemaking. First, artifacts are designed in order to shape practice in certain ways. The consequent effect on practice, however, is not a direct translation of artifact features to desired outcomes. Those who use artifacts perceive certain features as *affordances* (Gibson, 1986; Norman, 1993) that support a certain range of actions. Affordances are an actor's perception of the ways the artifact can be used in practice. The actual use of a complex artifact, such as a teacher evaluation policy, depends not only on the features built into the design of the artifact, but also on affordances of artifact use perceived by actors. The affordances perceived by local actors determine which features of the artifact are implemented. For example, an artifact that features evaluation in multiple domains of practice can afford a more comprehensive approach to teacher assessment by addressing out-of-classroom as well as classroom practice. The availability of these features does not mean the artifact will be used as intended. For example, an evaluator could focus only on classroom teaching behaviors while effectively ignoring out-of-classroom behaviors. In the hands of another evaluator, however, the evaluation artifact could afford a better-rounded assessment of professional practice. Artifacts can also serve to *constrain* behavior (Norman, 1993). Like affordances, constraints

are perceptions of artifact features that limit or qualify behaviors. Teacher union contracts, for example, often constrain evaluator action by permitting a maximum of two formal observation occasions during the school year. While certain affordances and constraints are built into artifacts by design, the challenge of implementation rests on the interests and abilities of local actors to identify and exploit the intended artifact features.

An artifact-based approach to the analysis of implementation focuses on how local leaders select certain features of complex artifacts as affordances and consider other features as constraints. Policy artifacts are introduced into schools not only to alter existing practices, but also to enhance the capacity of local actors to understand their work in new ways and to alter the organizational conditions of the work. A sensemaking perspective highlights how the introduction of complex artifacts draws upon and contributes to the evolution and interaction of individual understanding and local capacity.

School principals play a key role in how evaluation artifacts are implemented. In many school districts, administrative certification is required for performing teacher evaluations. Principals shoulder most of the burden of teacher evaluation processes. We use the concepts of principal will and skill and organizational structure to capture the interplay between actors and the school context.[3] The principal's capacity for innovation is measured in terms of individual *will* and *skill* to enact new practices. *Will* refers to the level of motivation of the local leader to implement the artifact. Leaders who have had a role or stake in the development of the artifact, and those who view instructional leadership as core to their role may be more likely to embrace the artifact, emphasizing its affordances and deemphasizing the constraints it may impose. *Skill* is the ability of leaders to engage in the intended practice. From a sensemaking perspective, skill levels are determined by the

relevant experience of the leader as well as by the training received for the intended practice. Will and skill are not generic capacities appropriately activated in predictable ways. From a sense making perspective, the availability of will and skill depends critically on how actors interpret the need for action in a given situation.

In addition to the will and skill of individuals, local leadership capacity is framed by the context of organizational *structures,* such as pre-existing practices and available resources, to support innovative practice. Leadership capacity is determined by the prior context of practice, including pre-existing similar practices, constraints on innovation, and multiple professional responsibilities. Our sensemaking perspective emphasizes how a leader's perception of structural possibilities, in the form of artifact affordances and constraints, bear on implementation. In the example of teacher evaluation offered above, the perceived needs and capacity of the local situation help to shape both a local leader's will to enact difficult features of a complex evaluation program, and her skill in fully implementing the artifact. Organizational capacity is both shown and determined through the material and temporal resources perceived necessary to support the implementation process.

Methodology

This study focuses on the ways that school leaders make sense of a complex district teacher evaluation artifact in their local school setting. We chose a case study approach to collect, interpret and present our data. Case studies provide opportunities to explore practices in depth, and to understand the complex interactions that characterize local systems (Stake, 1995). In order to make comparisons across cases, we chose to develop three cases of schools within a single district, faced with similar pressures to implement district policies.

Site selection

The study takes place in a large school district in the Western United States, which we refer to as Valle Verde Unified.[4] The district was chosen because it has made a substantial effort to implement a standards-based teacher evaluation system based on the framework for teaching (Danielson, 1996). The framework interested the district because it addressed criticisms of traditional teacher evaluation models by incorporating more sophisticated and elaborate evidence gathering and by providing feedback to enhance teaching practice for teachers at all skill levels and career stages.

We adopted a multi-dimensional approach to investigating how school leaders made sense of the teacher evaluation system. The data collected for the study include:

- interviews with district leaders and with principals and teachers from 7 elementary, 4 middle and 3 high schools in the district, for a total of 14 schools;
- written teacher evaluations in each school; and
- data describing the local demographic environment and instructional contexts.

Fourteen schools were selected from the district's elementary, middle and high schools. We consulted with a district representative to choose schools with a range of socioeconomic contexts and perceived levels of acceptance of the evaluation reform. Other schools were randomly sampled from among the remaining schools available.

Data Analysis

We began our analysis by examining themes that emerged from interviews conducted in all 14 schools. We searched for patterns in the how the artifact was used to support the principal's role as evaluator and instructional leader. The triangulation of principal self-reports with (a) teacher and district administrator interviews, (b) teacher assessment scores

and (c) written (narrative) teacher evaluations provide multiple sources of data to understand principal perceptions of the constraints and affordances presented by the evaluation framework.

Analysis of the data collected from the 14 schools shaped our selection of an elementary, middle and high school for more detailed case analysis of leadership sense making. We developed a coding scheme iteratively to allow patterns to emerge from the data. The coding scheme enabled us to explore the programmatic context, characteristics of the implementation process, local perceptions of artifact affordances and constraints, the impact of the evaluation system on principals, teachers and on the school, and local perceptions of artifact utility. After coding the data, we constructed three school cases to describe the implementation process in each school. The cases were then analyzed to reveal shared and unique characteristics of the sensemaking and implementation process.

Findings

In the following sections, we present a summary of findings from the teacher evaluation experiences in the 14 schools sampled in the district, including perceptions of administrators and teachers of system features and implementation. Following an analysis of the experiences across the schools, we provide illustrative case descriptions of the ways in which three Valle Verde schools implemented the new teacher evaluation framework. Each school case includes a brief demographic background, a description of the evaluator's perspective, an outline of the evaluation process, a summary of the written evaluation forms, and an account of the evaluators' and teachers' perceptions of the utility of the process.

Experiences Across the District

The evaluation system at Valle Verde, based on the *Framework for Teaching* (Danielson, 1996), was implemented in 2000 following three years of planning and field-testing. In contrast to the prior system, the new approach represented a more comprehensive set of teaching standards, with explicit performance rubrics, and multiple sources of evidence. The new district policy required that all teachers participate an evaluation cycle of a) a goal setting meeting, b) a pre-observation meeting, c) the observation, and d) discussion of the observation write-up. The cycle was organized around the district-developed evaluation model. The number of observations ranged from nine times per year for beginning, or probationary, teachers to single observations for experienced, or post-probationary, teachers.[1] Key findings relating to principal and teacher interview responses, written evaluations and evaluation decision-making in the 14 schools are summarized below:

Principal responses. Teachers and school leaders alike felt the evaluation system provided the opportunity to observe and reflect on teaching practice. Principal perceptions of the evaluation system ranged from an opportunity to develop morale or team building in the school to a significant time-management problem or a mandate that needed accommodation. Compared to the previous, open-ended system at Valle Verde, principals who viewed themselves as strong instructional leaders felt constrained by the specificity of the new system. Other principals liked the clarity of the new system for providing guidance on the focus of evaluation. Most principals viewed evaluation as a time management challenge, with increased meetings required and more paperwork requirements. Some made adjustments by streamlining their evaluation approach or cutting back on the amount and types of evaluation evidence. Others made changes to build in more time at school for evaluation activities.

[1] More details about the district evaluation policy are available in Appendix A.

Many gave up significant personal time to complete all of the evaluations. Each principal saw merits in the system despite the widespread belief that teacher evaluation itself was not a primary force improving teaching. Most evaluators adhered to the basic evaluation procedures and tried to complete the goal-setting session, the required number of observations, and the post-observation conferences.

Teacher responses. Teachers were largely positive about the feedback they received as a result of evaluation. With a few exceptions, feedback was seen as frequent, timely, and positive. Teachers cited specific examples of feedback that they utilized to change aspects of their instruction. Most said that their evaluator was qualified to provide feedback. However, in a few cases, teachers felt their evaluators were not adequately qualified to evaluate content-based pedagogy. In particular, evaluators who lacked instructional skills (e.g., those with a background in physical education, special education, or business) were not perceived as having the ability to evaluate instructional content decisions or pedagogical content knowledge. Few claimed dramatic change in instructional practice as a result of the evaluation process, but teachers were positive about the specific changes to theirs practice such as better questioning techniques, use of materials, and improved student engagement.

Overall, teachers were positive about interactions with their principals and other evaluators. Several post-probationary teachers remarked that their ability to select their evaluation domain contributed to the fairness, but not necessarily the accuracy of the evaluation. Others said that the principal or other evaluator set the stage for fairness by actively seeking dialogue with the teacher about the evaluation rating and getting the teachers' input. Several spoke of the principal encouraging teachers to offer other evidence if they disagreed with a rating.

Nature of evidence used in the evaluation. There was variation in the evidence gathered across evaluators. The evidence primarily consisted of class observations and related discussions. Although lesson plans and student artifacts were required to be collected, they did not appear to be systematically gathered or analyzed. In addition, some evaluators skipped the goal-setting session and either left out the goal-setting process or combined it with the post-observations conference (for the next series of observations).

The evaluation system was perceived as a low-stakes, formative artifact. Principals emphasized praise in written evaluations and provided 'gentle' criticisms if they criticized teachers at all. The district evaluation form contained an area for rating based on a number of rubrics and space for a narrative evaluation. Very little critical feedback was provided either through evaluation scores or in narratives. Principals did not assign an unsatisfactory rating in any of the 485 written evaluations we reviewed. For evaluation decisions, some principals evaluated teacher performance by comparing the teacher's practice to the proficient level (Level 2), and adjusted scores as evidence warranted. Others allowed scores to evolve more naturally from their analysis of the evidence. Narrative feedback was affirmative and seemed intended to foster reflection and growth. Written evaluations provided by elementary and middle school principals in many cases included longer narratives, despite often having more staff to evaluate than middle or high school principals. High school evaluations contained minimal written feedback, usually one to three sentences, even though the evaluation role was shared in high schools. Most evaluators allowed considerable teacher input into what would be observed and into the performance ratings (e.g., teachers could bring additional evidence to bear in the decision).

The analysis of the data from across the district revealed a substantial investment by district and local school leaders in designing and implementing the teacher evaluation framework. Many teachers and leaders were grateful for the opportunity to talk about their teaching. However, the reception of the artifact into local school contexts caused several conflicts. The evaluation program required a considerable amount of time. The time pressures, as we shall see in our cases, forced leaders to select which artifact features to implement. While the artifact was intended by designers to give local leaders a tool to improve teaching, most leaders did not use the artifact to disturb existing administrator-teacher relations. Praise rather than critique, and high scores rather than low, characterized the written feedback provided by evaluators. In the next section, we provide three cases to illustrate themes of how principals made sense of the artifact in their local school contexts.

La Esperanza Elementary School

La Esperanza Elementary is a K-6 school in the heart of the largest city in the Valle Verde district. Principal Susan Richards and her staff see the education of students learning English as a second language as the main challenge for the school. La Esperanza's 36 teachers are organized into grade-level teams throughout the school. 81 percent of the 690 students are members of a minority group (primarily Latino), and 84 percent of the students qualify for free and reduced lunch. Nearly half are classified as English as a Second Language (ESL) students. The school is currently under significant accountability pressure from the state and is being monitored and assisted in the effort to improve student academic performance.

The teachers interviewed at *La Esperanza* included one first grade, one second grade, and two third grade teachers. Three of the four teachers did not have substantial teaching

experience, while the fourth had been teaching for more than a decade at the school. All four of these teachers (along with the rest of the faculty) were evaluated by the principal. All of the teachers interviewed and the principal agreed that *La Esperanza* had challenges not faced by other district schools. Principally, the presence of significant numbers of non-English speaking students meant that teachers must be patient with and accommodate students.

Evaluator characteristics. Principal Richards had been at the school for four years. Before coming to *La Esperanza* elementary, she worked for eight years as a fourth grade teacher, served as a teacher leader, and a trainer for the district initiatives in writing and math. In addition, she worked as a dean of students for four and a half years and as an elementary school principal for three years. Part of her prior work was on a Native American reservation, working with a highly at-risk student population.

Richards viewed her role as an instructional coach for the faculty. She believed that her experience with the district provided her with the knowledge and skills she needs to identify appropriate teaching techniques and make helpful suggestions. She recognized the potential stress associated with a summative evaluation system that attempts to provide formative feedback to teachers, and works with teachers to reassure them that the system is formative and an opportunity for growth, rather than for humiliation and anger:

> My goal is to make them, to help them feel more comfortable, that I am not just an evaluator but I am also a coach. That is my role. That is the role I want. I want to be a supervisory coach. And so we work hard at trying to establish that kind of rapport. And we are getting there. It has taken four years of trust to know that I am not going to beat them up…and destroy them.

Teacher interviews corroborated Richards' description. All four teachers commented on their positive and upbeat interactions with Richards. One teacher said that "she truly is there to help us. I mean, not to criticize or anything like that…I love it when she would come in because I know she is watching me to help me improve what I am doing."

Evaluation process. Richards estimated that she spent approximately fifteen hours per year on each teacher's evaluation. (With 36 teachers, this is the equivalent of fully a third of the academic year spent on observation, evaluation, and feedback). The evaluations were based on evidence collected through formal observations and intermittent informal observations, such as walk-throughs, throughout the year. The principal also gathered information regarding teacher performance during other committee meetings and professional gatherings. She viewed the new evaluation system as flexible in its use. For example, this past year she chose to emphasize teacher goal-setting and required all teachers to submit their goals in the first month of the school year. As part of goal-setting sessions, teachers evaluated themselves and then discussed her evaluation of their performance. Richards connected the evaluation process to how teachers met their goals.

In her classroom observations, Richards split her time between scripting part of the lesson and observing classroom dynamics. This process allowed her to get a sense of how the classroom worked while using examples of classroom conversation and activity in her report. She focused on teacher skills in questioning and responding to students. After recording her reflections on the observation form, Richards dropped the written evaluation off for the teacher to sign and arranged for a post-observation conference. During the conference, she asked about the strengths and weaknesses of the lesson and what the teachers might do differently. She offered positive feedback to highlighting the successful aspects of the lesson.

Richards reminded teachers of needed changes in a positive manner "until the third time, that is my key, a third time…if I have asked you three times to clarify your lesson plans …and they are still not clarified, then it becomes an evaluative measure... that is all lettered and documented."

Summary of written evaluations. An analysis of the evaluations revealed that the mean score for the faculty across evaluation domains at La Esperanza was between "proficient" and "area of strength" on the district scale. No teacher received an unsatisfactory rating. This indicated that, according to the principal, most teachers are performing at or above a proficient level.

Richards included a significant number of written comments on the teacher evaluation forms. The narrative section averaged just over 24 sentences per evaluation. The majority of the narratives were composed of excerpts from Richards' scripted observation notes. Each evaluation had a final summative paragraph, which expressed the high value that person added to the school. In addition, this final paragraph always included a sentence saying that the teacher was an important member to the *La Esperanza* family. For example, she described one teacher as being a "wonderful asset for La Esperanza."

There was no clear relationship between the assigned scores and the amount and content of the written narratives. For example, there were several instances where a score of a 1 was given, but there was little or no discussion of the rationale for the low score in the narrative. Information gathered from the teacher and principal interviews suggested that substantial dialogue was taking place between the principal and the teachers that was not documented in the evaluation forms. For example, one teacher reported that, after an observation, the principal gave her a recommendation about how to improve her questioning

and answering techniques with her students. This recommendation could not be found in the written evaluation narrative. The teachers also reported that they regularly received verbal suggestions from the principal throughout the school year.

Perceptions of the evaluation process. All four of the teachers interviewed indicated that Richards' written evaluations are extremely affirmative, emphasizing many positive aspects of their teaching. As the principal said, "I try to find their highlights…and then I will make one recommendation. I won't beat a dead horse, but I will make one recommendation because that is what I should do and to assist my teachers." The principal indicated that teachers have been positive about the evaluation process. She gave the following example of a positive response from teachers as a result of the evaluation process:

> I was in a first grade classroom and I saw a lot of the same writings hung up on a wall that had been there all year. And I didn't have. . . a problem with that but I was curious for the teacher to tell me what was the purpose to maintain those? And their reasoning was excellent. Based on this reading training that we had which is the children go around and they read familiar print continuously so that they have success with finding the writings and it is also finding words and they do word writing. So that kind of conversation is really good so I understand what they are doing. And they also are aware of what is going on in the classroom.

The principal's focus on positive feedback and coaching, along with the team-based instruction throughout the school, helped to create a climate of openness to observation, evaluation, and feedback among teachers. Teamed teachers often worked together to address evaluation goals. While the evaluation was not a primary focus, teamed teachers typically discussed with one another their evaluation goals as they planned their work for the year.

Despite the large time commitment and teacher reception to the evaluation system, Richards did not believe that the evaluation process was a good tool to improve teaching in her school. When asked whether the evaluation system could change teaching, she said,

> On average, no. I don't think so. … I think it can be very disheartening. I think evaluations can either encourage and give teachers a pat on the back that they don't often get or it can totally destroy them. It is just how you approach it. And I have seen both things happen. It is a very hard thing to do. And I don't think it changes people. I think it can stop people. I don't think it changes them.

While the evaluation system itself may not have led to deep change, the principal described how state accountability requirements provided pressure to change.

> I think what changes us, what drives change here for my teachers will be, well, it really comes down from the State Department beating us up. And then as a team, it is a total team effort, we get together and look at our scores, look and what we are doing, and then we look at what we need to change and how to implement change?

Richards' low estimate of the impact of the evaluation system contrasted with teacher's views. Teachers provided several examples of how Richards' evaluation feedback enhanced their teaching. Newer teachers remarked how the evaluations helped to improve classroom management. A veteran teacher offered an example of feedback regarding pedagogical content in math. Much of this feedback was specific and not directly connected to the evaluation framework. For example, one teacher indicated that the principal suggested using a microphone at the next student presentation and having a master of ceremonies to host the show. Another recommendation focused on increasing wait time after questioning

students, or shifting the balance of large and small group time to improve student discussion. Teachers did not mention evaluation rubrics in their comments about principal feedback.

Several new teachers believed that the framework itself provided a "progress map" to identify areas for improvement. One teacher said that the system "kind of gives you the direction to go to or work towards." While these new teachers noted the effect of the system on their practice, one veteran teacher did not think that the system caused teacher change. However, the veteran teacher did say that the system provided a good "method to track" what type of professional development she would seek. All four teachers believed that the system was fair. The teachers reported considerable input into what went into the final evaluation. One teacher said that she can "discuss with her why I feel I am at a 2 and maybe at a 1 there. She is very fair about taking my suggestions into her reasoning."

Woods Middle School

Woods Middle School serves about 1000 6th, 7th and 8th graders in a large city in the Valle Verde district. In 2001-02, 17% of Woods students were Latino, and 28% qualified for free or reduced-price lunch. Student performance in reading, language arts, and mathematics is above national norms for eighth grade students; however, significant gaps exist between the test scores of white and ESL students. The school staff included 42 teachers, five special education teachers, and five additional teaching staff. The teachers interviewed at Woods included one probationary teacher, two teachers on major evaluations, and two on minor evaluations. The probationary teacher was in his first year, the two major evaluation teachers had been teaching for less than five years, and the two minor evaluation teachers had been at Woods for at least eight years. Two taught math and three taught English.

The school was organized around a cohort model that grouped students and teachers together as they passed through grade levels. This structure gave teachers opportunities to get to know their students well, and established structures for common instructional planning time. While several of the veteran teachers mentioned this structure as an occasion to share strategies about instruction, other teachers designed and taught their lesson plans independently.

Evaluator characteristics. The Woods administrative team included Principal John Storm, an assistant principal and a Dean of Students. Storm was in his seventh year at Woods Middle School, his third year as principal. After an earlier career as a managing partner of a private sector business, Storm has spent his past twelve years as an educator in the district. Storm feels that his main strength as a principal has been his ability to listen to teachers, students, and parents and to solve problems as they emerge in the school. This blend of problem-solving and listening has enabled him to use the evaluation system to point out potential instructional issues while being sensitive to teacher's professional context:

> Because then I can point out problems to (the teachers) …and that requires a little bit of discussion. Some of it just comes from personal experience. I have been doing this long enough where I happen to know that so and so is working on their master's degree. I don't have to ask them. I just know they are doing it.

Principal Storm saw the teacher evaluation framework as an important, if burdensome, supplement to his role as school instructional leader. Storm felt the evaluation system was particularly useful as a tool to help or dismiss probationary teachers. His approach to the new evaluation program was informed by his own six-point system for what constitutes good teaching:

> The first is that the objective is clearly stated … and the kids have to know what it is they are supposed to learn. The lesson has to have a clearly defined structure. You can't just do this and that. They all have to be related. I expect to see most of the students actively participating in their learning, not just sitting there and listening. I expect to see a teacher checking for understanding frequently so that they don't keep teaching after they have lost their kids. And then I expect to see teacher/student, student/student interactions to be appropriate. And any misbehavior I expect the teacher to respond to appropriately.

Teachers reported that Storm's six-point system characterized their experience of the evaluation process. Four of the five teachers interviewed reported that the principal's concerns with checking for understanding, classroom management, and clearly defined organization structure came across in the evaluation process. When asked to provide a specific example of feedback, four of the teachers mentioned Storm's review of their questioning practices. The teachers did not seem to differentiate between Storm's established checklist and the new framework. One veteran teacher noted that Storm's focus on questioning technique flowed from "the evaluation form he always uses."

Evaluation process. Storm used the new evaluation system as a complement to his existing informal system of formative feedback. Storm began the evaluation process with brief visits to each classroom within the first several weeks of school. "Leading up to that I spent some time in the classroom informally, two times that I documented, and then two or three times just walking around getting into the classroom." His multiple observation practice enabled him to get a sense of where potential problems might occur in the school as well as to introduce himself to students and teachers throughout the school.

Storm considered the observations themselves to be an important component of the evaluation process. Allocating sufficient time to observe all teachers requires annual planning. "What I do is I reserve 25 percent of my day, one period a day, to do observations. And then I will sit down with the teacher's schedule…and I will actually book observations a month in advance." His time commitment to the evaluation process – fully 25% of his time -- is corroborated by the 114 formal visits recorded on the official evaluation forms.

The annual cycle began with an opportunity for teachers to rate themselves using the evaluation system:

> At the beginning of the year I ask the teachers to go through the rubric and self-evaluate. And then when they sit down with me to go through their goals for the year, I will ask them about their self-evaluation….I find teachers to be pretty much on target. They know where they are.

Storm then scheduled individual teacher observations. He used a laptop to record his observations of classroom practice. Storm's ability to write-up his comments in the class enabled him to provide feedback to teachers by the end of the school day. These comments serve as a rough draft for the final evaluation report, and give teachers the chance to discuss the main points of the report before it takes final form.

Storm relied on his past experience as an evaluator as well as the observation data to make his judgment about the quality of teaching. Storm began each rating at Level Two, the basic level of performance. If the teacher has met the Level Two criterion, Storm moves to Level Three. "Level Three is just a little extension of Level Two. In fact, most of the rubric is written, take Level Two, and then add a little component to it." Level One ratings provide a special challenge in writing the final report. Storm commented: "Level One is poor teaching

even though it is satisfactory, it is still poor teaching." Storm felt that the level of documentation must be much greater in a Level One evaluation, as it is directed toward remediation or to establish grounds for termination. Consequently, Storm reported that teachers with Level One evaluations received more substantive feedback for their observations.

Summary of written evaluations. 48 Woods teachers were evaluated during the 2001-2002 school year. Average scores ranged between "proficient" and "area of strength" across the four evaluation domains. No teachers received an unsatisfactory rating in any domain. Limited narrative feedback was provided for each teacher. Forms for post-probationary teachers included an average of 9.7 sentences per domain area, while the probationary teachers received 10.2 per area. Most of the narratives included a balance of descriptive and laudatory sentences. There was an average of less than one sentence per evaluation directed toward either suggestions or critiques of teaching. Although both the teachers and the evaluators remarked on the value of the scripted comments made in class, these scripted comments were not present in the written evaluation forms. Over half of the evaluations included sentences commending the teacher's contribution to the local school culture, hard work, or participation in extra-curricular activities.

Perceptions of the evaluation process. Teachers were generally more positive than Storm about the potential effect of the evaluation process on their teaching. One teacher commented how the rubrics and domain structure of the evaluation program "helps create a common sense of good teaching" among the staff. Another teacher mentioned that the framework offers a structured opportunity to reflect on practice that "helps me strengthen my content knowledge." Teachers differed about their assessment of the Storm's time

investment. Two teachers noted that, even though the time taken by the observation and evaluation process signified administrative interest in teaching, the principal did not spend enough time in their classroom to really make a difference.

Even though he questioned the effect of the new system on shared perceptions of teaching, Principal Storm saw the teacher evaluation program as an improvement on the system it replaced. The older system focused on nine "topics" of teaching, and allowed teachers to pick three topics on a major evaluation, and one topic for a minor. The disadvantage of that system was that it allowed teachers to "focus on one area and just ignore everything else." The new system:

> Forces you to look at a broad range of teacher skills. And in that respect it is very, very good. Because, as an administrator, I am looking at this, boy, they have a lot of stuff they have to do as teachers. And it helps me remember the things that I am supposed to be looking for.

In Storm's view, the new system was particularly helpful in documenting poor performance and for helping new teachers. These affordances of the system accorded with Storm's belief in the importance of working with probationary teachers. The system rubrics provided a common reference for communicating about substandard teaching practice. Storm offered an example of how:

> In Domain Three, under grouping of students, if I were to tell a teacher that his or her instructional groups are inappropriate to the students or the instructional goals, that is unsatisfactory. Now, if a teacher knew that, then they could go to this rubric and say, well, what is satisfactory? … So, for someone who is doing poorly, it is very beneficial."

The system helped Storm and the teachers frame formative programs to improve their teaching. Storm described a teacher evaluated as unsatisfactory several years before: "I was very, very specific in the areas and had very concrete evidence as to why he was unsatisfactory. And I have worked with him for four years now and I would say this year he has made some real improvements."

According to Storm, the capacity of the system for identifying and helping poor teaching did not seem to apply equally well to good teaching. Because he spent the most time with newer or poorer performing teachers, good teaching received relatively less feedback. Storm contrasted the value of the system for probationary and proficient teachers:

> But I view this system as extremely effective for an unsatisfactory or a Level One teacher. For a teacher who is proficient or a very strong teacher, we are just documenting the fact that they are good teachers.

Storm did not feel that the new evaluation system supported the establishment of agreement about what good teaching means: "I think every teacher thinks that their teaching is good. Whatever they do is good. And they haven't tailored their teaching style to meet the rubric." The novelty of the system may mean that it has not had a chance to create a shared sense of agreement. Storm commented, "you have to remember . . . these teachers weren't brought up in this system. This system has been imposed on teachers that have been here a long, long time." It is interesting to note that while Principal Storm emphasized the value of the evaluation process for novice and poor teachers, it was the veteran teachers with relatively higher scores who reported the most benefit to their teaching. Two veteran teachers valued the opportunity to reflect on their teaching afforded by the process. One teacher mentioned

that the "rubrics helped me understand the difference between Level 2 and Level 3 teaching," and that the rubrics gave him something to aim toward in his teaching.

Storm's final point concerned the lack of feedback he has received on being an evaluator. While he noted that he received valuable training to conduct evaluations, he has received no feedback on his own evaluation practices. In his words:

> I have never gotten any feedback from anyone on [whether] I am doing a good job as an evaluator…. I mean, my bosses never ever talk to me about the evaluations that I have written. I don't think my boss has ever read my evaluations. So, I wouldn't mind getting some feedback to know whether or not I am meeting district standard or not.

Jaye High School

The Jaye High School context presents special challenges for understanding how evaluators make sense of the evaluation process. In 2001-02 the largely upper-middle class student population at Jaye included 1,880 students. The student transience rate was 16 percent, 6.8 percent were labeled as special education students, 3.1 percent of the students were English-language learners, and the free/reduced price lunch population was 11.2 percent. Student performance on a national norm-referenced test was higher than the average performance of other district high schools. The 93 teachers on staff included 10 probationary teachers. The teachers and principal commonly reflected upon two features of the school context during the interviews. The first involved efforts to involve staff across the curriculum in setting school goals. The second, and a related factor, was a strong sense of collegiality in the school.

Teachers interviewed at Jaye included a veteran English teacher, a mid-career biology teacher, an early/mid-career history teacher, and a novice mathematics teacher. Principal

Jennifer Fredericks was in her third year at the school and had extensive experience as a teacher and administrator. The four other administrators who acted as evaluators were not interviewed.

Evaluator characteristics. Fredericks sought to develop department and school-based instructional goals, using a consensus-building approach. She explained that from her first day in the school, she worked to get the school to focus on data (e.g., student test scores) to set goals and to monitor progress. She encouraged teachers to share best practices during staff meetings. These processes were intended to develop a "common building belief system of what we are doing as a community, the sense of community to serve the students."

Principal Fredericks supported the new teacher evaluation system and was willing to invest the time and effort to make it productive. She asserted that the system fit with their "school wide belief system in rubrics. … It gives you a verbal picture of what you want to see." Her active support of the evaluation system capitalized on her instructional expertise and was reflected in how she structured the evaluation process. Fredericks described the evaluation system as fitting her philosophy on instructional leadership and incorporated the rubrics into her "own contextual belief system." She compared her leadership approach to the four domains of the evaluation system by planning what she wanted to do before she became a principal (reflecting domain 1); creating an environment "where people felt free to interact with me, to interact with one another," (modeling domain 2); then implementing the plan or plans (domain 3); and finally, giving back to her school community by working and sharing with each other through best practices during faculty meetings (domain 4). As she summarized, "We have actually role-modeled the [evaluation system]" as school leaders.

Fredericks explained that her experience, training, and practice as a teacher made her a good evaluator. Before she became an administrator, she attended Madeline Hunter training sessions and developed her skills in scripting classroom observations. In addition to the skills needed to conduct evaluations, Fredericks asserted that her credibility as a classroom teacher was a critical attribute for her legitimacy as an evaluator. As a principal, she saw her most important strength as her "ability to see all sides of the situation and to put myself in the shoes of the other person, whether it be a parent, a student, a teacher; to understand where everyone is coming from. And not to take anything personally."

Despite the increased demands of the evaluation system, Fredericks said she was able to manage the process because, "I am a pretty good time manager … I look over a semester and I can figure out where I have to be." To handle the time and workload demands, she planned one semester at a time and began with the probationary teachers, who required more time due to the structure of the evaluation system and their uncertainty in practice. Then she worked with teachers on the major evaluation and finally addressed the minor evaluations, "… because they take less time." She lamented that the time dedicated to probationary teachers, although necessary, limited how much she could work with other teachers.

Evaluation process. The evaluation process described by the principal was similar to that described by teachers who had other evaluators. Fredericks held pre-observation conferences to meet with the teachers before the evaluation process began. During the meetings, she went through the evaluation rubrics and procedures to explain the process. She asked where teachers saw themselves in the rubrics. If a specific rubric lacked clarity, she would discuss what she believed it was trying to get at. When the questions were resolved, the focus of the evaluation was selected (i.e., domain(s), components, and elements).

Teachers set a target for growth for each domain. Teachers chose a growth goal for each domains on which they were evaluated. Probationary teachers prepared one goal for each of the four domains. The targets of growth served as the focus for the written evaluation.

Fredericks structured her evaluation approach to focus on probationary teachers and centered her efforts on maximizing formative feedback to these novice teachers. She assigned herself a larger share of probationary teachers than the other evaluators. As she explained, "I want the new teachers I hire to have that connection with me and…I think that is a very formative time." She saw probationary teachers as vulnerable and was concerned about attrition, because new teachers typically "are just not mentored and encouraged."

To lower teacher anxiety about the process, Fredericks told teachers to feel free to make mistakes and not to worry about her being in the room. She also gave them flexibility in scheduling observations. As she stated, "I am not there to look at a perfect lesson … so I try to put them at ease ... because I see this role as a helping role, not to go in there and catch them doing something wrong." She also tried to make sure that teachers were aware that what was being written down would be in the evaluation. As she explained, "There is never anything in the evaluation that I do that surprises. There is never anything in writing or checks in those boxes that the teacher has not been with me [and discussed]… And I try to always find something to commend them on." She explained that she was careful about what she wrote and how she phrased written comments in order to prevent teachers from reacting negatively to evaluations. When she first started doing evaluations as an administrator, she "… was amazed that the use of a word could make somebody very anxious." So, "… I am more careful about using words like 'very' or 'often' or 'frequently' or 'occasionally'." During conferences, she asked teachers to talk about instructional artifacts (planning

documents, test results, etc.) involved in the lesson. Consistent with the leadership approach discussed above, teamwork and school wide goals re-emerged during evaluation discussions.

Fredericks tried to foster self-reflection and monitoring/correcting and used constructive criticism, trying to help teachers think about the observed situation "… and go back to it in their mind and think about how they might do it differently. And then I will say 'or you could have …,' but I don't say 'this is the way it should always be done.' There is never one way to do anything." She tried to encourage teachers to have interactive classes, where kids are major participants.

Summary of written evaluations. The evaluation context at Jaye is made more complex because of the multiple evaluators involved. Thus, even though Fredericks played an important role in making sense of the evaluation artifact for the school program, the other evaluators brought their own assumptions to the process. Thus it is not surprising that both the written evaluations by evaluators and the teacher reactions to their evaluations at Jaye varied considerably. Two of the evaluators (including the principal) provided detailed written commentary, with evidence described and specific recommendations for improvement. In contrast, the other evaluators provided very brief descriptions of performance, with only a few sentences, little if any evidence reported and few recommendations for improvement.

Individual teacher evaluation scores on the 79 evaluations provided by the district averaged between "proficient" and "area of strength" on the district scale. Although five teachers received level one ratings in particular domains, there was no written description of why the rating was given or how the teacher could improve on the element. Despite the prompt on the evaluation form (and implied requirement of the system) to offer specific evidence for the ratings, it was rare for evaluators to offer such evidence or to provide

recommendations to improve. It was also difficult to find negative feedback in either the write-ups or from the interview transcripts. Evaluators delivered criticism in a positive fashion (if critique was provided) or first pointed out positive aspects of performance and characteristics of the teacher.

Written evaluations documented positive aspects of teacher performance. In some cases, recommendations for improvement were provided. For example, one evaluator commented that, "teacher uses goals suitable for most students." This same evaluator had three recommendations for the teacher, including the following: "When planning lectures, provide as many opportunities to engage as many students as possible throughout the lecture." The written evaluations also allowed evaluators to document praise for how teachers had taken on extra school responsibilities. For several of the teachers interviewed, more feedback seemed to be provided during discussions with the evaluator than was reflected in the written evaluations. However, other teachers reported receiving minimal feedback in either written or verbal form.

Perceptions of the evaluation process. Principal Fredericks thought the comprehensive standards and rubrics of the evaluation system helped to promote a common and continuing dialog with teachers. Fredericks believed the system provided a framework for teachers to think about their work and a process for them to interact, get help and talk about their practice, and be recognized for their efforts. Most teachers also preferred the new system to the prior one, which required an extensive written evaluation but did not have the level of knowledge and skill elaboration of the current system. One teacher, however, preferred the old system, where "… you could sit down and talk and you could read it and

pick it out and read it again if you need a little pat on the back." Other teachers valued the potential for objectivity of the new system's detailed rubrics.

Jaye staff had mixed reactions about the impact of the evaluation process on their instruction. The principal believed that the evaluation system led some teachers to change their practice. For example, a special education teacher who had relied on lectures changed his practice to get students more actively involved. After evaluation discussions, Fredericks noted "… his room is full of colored pens and pencils and the kids have no books and the kids are keeping these forms… And he loves it. " Two teachers mentioned the evaluation process improved their teaching through better planning and classroom management, keeping students on task and increased use of reflection. Two other teachers were not as positive. One commented that her evaluator (not the principal) had little or no teaching experience: "I was evaluated by someone who didn't teach school, who has never taught school. [He] went from the world of work, business, into education, into administration." This teacher reported a better experience with a different evaluator the prior year. The other teacher consistently received high ratings and was rarely offered feedback specific to the content he taught. Teachers commented that the evaluation system required more paperwork and effort than the prior system, but it was more burdensome for evaluators than for teachers. One teacher said that, "I think what happens is [the administrators] get up against the time when all the evaluations are due and things get really hectic." Two teachers explained the system was more work intensive than the prior system, but was worth the effort and more objective.

Fredericks expressed that evaluator training offered by the district could be improved. The trainers took a minimalist approach focused on getting evaluations done efficiently rather than well. As she stated, "Basically, the person was saying, 'this is how you can get

them done the fastest. You don't really have to do this. And you don't really have to do that. And you can just whip them out." She suspected that evaluators varied considerably in how accurately they evaluated teachers and how well they provided growth-directed feedback. She suggested the district should have master evaluators help beginning administrators, to "go in until they have a comfort zone of performing evaluations and the process."

Analysis

Investigating the way that local leaders make sense of a complex artifact such as a new teacher evaluation highlights the selection of artifact affordances, from among the many possible features of the artifact, and helps us to understand how leaders adapt new practices to their existing contexts. The interaction of leaders' will, skill and their perceptions of organization structure organize our comments about sensemaking.

Will

Most principals wanted to make this system work and tried hard to comply with the system requirements for numbers of observations and write-ups. However, it was apparent that the evaluation system was extremely time-consuming, absorbing as much as 25% of the principal's time. We saw principals address the time issue by complex scheduling and by investing significant amounts of personal time. Evaluators satisficed the time requirements through brief classroom visits, writing up the observation while observing the class, and in stealing a few minutes before and after class for the pre- and post-observation meetings. Despite this significant time investment, some teachers felt that an insufficient amount of time was invested in the system to provide meaningful feedback on teaching practice. In many schools, most evaluations were dated all on the same day at the end of the evaluation cycle, suggesting that many evaluation forms were completed at the last minute.

The considerable time investment required to conduct observations and complete the evaluations narrowed the range of cognitive and structural resources available to implement the full range of artifact features. The artifact design relies on evaluators to collect multiple kinds of evidence to document the different components of teaching practice as specified in the rubrics. The elaborate system of rubrics and evidence requirements challenged evaluators to move beyond classroom observation in order to develop fundamentally new evidentiary bases. Our study suggests that although evaluators stretched their professional and personal time to observe all teachers, the evaluations lacked evidence grounded in the rubrics. The evaluation criteria included, for example, "reflecting on teaching" and "communicating with families," but no evidence was provided by evaluators for ratings in these domains. Simply complying with the district policy to conduct observations of all faculty seemed challenging enough. To take full advantage of the evaluation program, evaluators and teachers need more time and training on how to collect, reflect upon, and present evidence to maximize the potential of the evaluation system for promoting better teaching practice.

Skill

We found that the written evaluations lacked either formative or critical feedback. The majority of written comments focused on scripting of classroom activities, classroom management and generic comments pointing to the important role the teacher had played in the school. While several of the principals used the evaluation process to suggest new practices and to encourage staff collaboration, few examples of specific, evidence-based suggestions grounded in the rubrics found their way into the written evaluations. The focus on classroom management was reflected in the views of new teachers who expressed a more positive view of the potential impact of the system on improving their teaching practice.

Veteran teachers, presumably more familiar with classroom management practices, were more reserved in their praise.

The evidence from our case study schools suggest that evaluators lacked the skills to provide valuable feedback, particularly with accomplished teachers. Evaluators instead used evaluation as an opportunity to work with novice teachers and to build a positive school culture rather than as an opportunity to push instructional practices to the highest levels. However, we cannot discern from our study whether this lack of skill was a cause or an effect of evaluator priorities. In other words, the perceived lack of skill in providing formative feedback to accomplished teachers was qualified by the competing, and perhaps more legitimate, goal of enlisting the support of veteran teachers to the new evaluation initiative. Concerns about the politics of evaluation and maintenance of strong social relations among faculty and evaluators may have led evaluators to provide nearly exclusively positive and largely low level, narrow and specific feedback to teachers.

The lack of critical comments and the inconsistencies between the reported value of feedback and the written instruments suggest the importance of attending to the political context for evaluation. While the lack of "unsatisfactory" ratings in the case-study schools and the narrative feedback might suggest a high quality of teaching across the schools, all principals described instances of sub-standard teacher performance. Clearly, the evaluation process was not fully represented by the written components alone. Performance appraisal research suggests that negative feedback is difficult to convey and often avoided for fear of depressing employee motivation (Ilgen & Davis, 2000), and the political nature of formal appraisals may result in lenient evaluation ratings in order to motivate employee performance (Longenecker, Sims & Gioia, 1987; Murphy & Cleveland, 1995). In such cases, evaluation

systems may send mixed messages about organizational goals for rating accuracy and performance improvement through evaluations (Kozlowski, Chao & Morrison, 1999). Written negative comments carry great weight in organizational cultures, and supervisors interested in maintaining long-standing, collaborative relationships with employees are often reluctant to use formal instruments to provide negative feedback.

The absence of critical feedback in most written evaluations might not mean the complete absence of such feedback. Recall Principal Storm's comment that the specificity of the rubrics was valuable for helping to dismiss incompetent staff, yet these critical messages were not reflected in the written evaluations. If teachers receive the most meaningful feedback verbally, then the written instruments could be used to preserve the delicate organizational culture of trust and collaboration between evaluator and teacher. At the same time, neglecting to document specific instances of low performance blunts a central intention of the evaluation program. Without critical feedback, the artifact becomes a tool to maintain a positive sense of community rather than a tool to distinguish levels of practice, and to foster improvement and reflection on teaching practice.

Structure

Structure here refers to the personal, professional and institutional traditions that shape local practice. Our analysis showed the power of the self-perceived role of evaluators as instructional leaders on the evaluation process. Self-imposed role definitions reflected the skills of the evaluator, and seemed to enhance or constrain their will for selecting and implementing certain features of the artifact. The roles chosen by evaluators had significant effect on the affordances of the artifact selected for implementation. For example, Jaye's Principal Fredericks actively modeled her instructional leadership approach around the vision

of performance represented in the evaluation domains. At *La Esperanza*, Richards' role as an instructional coach was reflected in her team-building messages of encouragement and inspiration on her written narrative evaluations. Richards' belief that critical evaluation feedback could be devastating to teachers shaped her role as an evaluator to encourage rather than to criticize her teachers. She downplayed the summative, critical features of the artifact in order to fit the artifact to her perceived role in the school.

Woods Principal John Storm perceived the evaluation artifact differently. His role as an instructional leader involved communicating a consistent message about his six key indicators of good teaching. While not inconsistent with the evaluation model, it was these indicators – and not the evaluation system itself - that guided his observations. Storm relied on his model to guide the Woods evaluation process and to give specific feedback to struggling teachers. In this case Storm replaced designed features with his own conception of good teaching, and used the new district initiative to flesh out his previously developed evaluation practices. While full implementation of the artifact may require a redefinition the self-perceived role of the evaluator, the expertise of the evaluator as an instructional leader depends on the very role-perception in need of alteration. Implementing more of the artifact features would require evaluators to "see" their instructional leadership roles differently to allow for a more critical perspective on evaluation practice.

Conclusions and Implications

In this study of sensemaking and implementation of a knowledge and skills-based teacher evaluation system, we found that the features of the artifact that potentially enhance the opportunity to improve teacher quality were filtered through pre-existing perceptions, knowledge, and structures. Consistent with the literature on sensemaking and

implementation, we found that local implementation of the evaluation system varied substantially from school to school, and was shaped by the ways in which principals understood their own role, their context, and the evaluation artifact. Principal sensemaking seemed to be primarily a function of principal self-perception of their role as a leader and the knowledge and skills they bring to that role; prior evaluation practices in the school and district; and school context factors such as teacher morale and existing challenges facing the school (e.g., student population risk factors, external accountability pressures).

We found that there was a strong desire of local leaders to use teacher evaluation practices for two central purposes: one, to maintain a community of good will with teachers, and two, to help novice teachers improve or remove those unable to perform at a basic level. In each case, the affordances exploited by leaders seemed to extend the functions of the previous evaluation system. Further, these uses seemed to inhibit the recognition and use of other features intended to provide specific, critical, and formative feedback to veteran teachers.

A key question in the implementation of complex artifacts is whether features have sufficient power to change the embedded organizational culture. From a compliance perspective, the amount of time spent to implement the teacher evaluation framework should be judged a huge success at Valle Verde. However, implementation of the full range of artifact features seems hindered by time constraints and school cultures and professional practices that reinforce the separation of instructional and supervisory practices. The gap between supervision and instruction that constitutes the organizational culture of many schools is difficult to cross (Rowan, 1990; Hazi 1994). Closing this gap takes time. A condition for closing this gap might be to develop both common practices for teachers and

leaders to interact around instruction, and a common language to facilitate the conversations. The district framework for teaching includes features to facilitate both processes. Since the framework is already in place at Valle Verde, and since teachers and principals view it as a useful process, there already is significant movement toward these ends. The framework is being used widely across the district, and appears to be helping to develop the capacity for teachers and evaluators to engage in regular conversations about instruction. District leaders could push implementation further and capitalize on this newfound capacity in order to more tightly couple instructional and supervision practices in the school culture. Over time, this capacity may have the power to change instructional culture.

Thinking of implementation as a long-term process of reshaping prior knowledge, skills, and beliefs will require district leaders to focus on the key "teachable moments" currently emerging for district evaluators and teachers, such as the desire of evaluators to receive district feedback on their own evaluation practice. We hypothesize that the increasing experience with evaluation and feedback might make principals more likely to identify and focus on the instructional improvement features of the evaluation system. Taking advantage of the ways evaluators learn from their experience may change the features principals select in the artifact, and therefore modify the implementation of specific features of the artifact to enhance its instructional improvement outcomes. Specifically, our analysis suggests the following five areas of focus for continued attention:

Providing a Clearer Conceptual Connection between the Teacher Evaluation Framework and Enhanced Student Learning

A key intention in the district design was to use the evaluation artifact to improve student learning. However, few principals and teachers viewed the evaluation process as

having a direct relationship to student achievement, accountability goals, or even as a pathway to significantly improving teacher quality. To make this link explicit, evaluators may need additional training in content-based pedagogy and evaluation feedback. Enhanced skills alone may help evaluators recognize the opportunity for instructional improvement that the evaluation artifact provides, and thereby encourage them to use evaluation as a means to work with teachers at all skill levels to significantly improve instructional practice.

Tying Feedback to the Evaluation Standards

Training for evaluators could also focus on building understandings about how evaluation rubrics enhance teaching practices and improve student learning. Although teachers are asked to set goals for one or two specific elements in the domains on which they are evaluated, written feedback is seldom specifically tied to the standards. Maintaining a focus on the evaluation standards beyond the goal setting process could help to more directly link goal-setting, evaluation feedback, and overall improvement in the teacher evaluation system. Training in providing evidence-based feedback, such as evidence needed to demonstrate content-specific pedagogy, could extend existing training and support relationships in order to create shared understandings of evaluation as a tool to promote instructional improvement. Teachers could more clearly see a connection between formative recommendations and improvement on the rubrics in the evaluation system.

Coordinating the Structural Requirements of the Program

With three years of implementation, the routinization of the evaluation process provides a foundation for further development. District leaders need to familiarize themselves with the evaluation process, and better understand the various roles that goal setting, observation, and verbal and written feedback play. In doing so, the district could

provide feedback on the existing evaluation process in local schools, and evaluators could develop networks to share practices on how to provide effective and efficient evaluations. Once district and school leaders realize how far they have come, their insights can be used to build on these newly developed capacities.

Recognizing and Accommodating the Political Contexts of Evaluation

The politics of evaluation were evident in a variety of features of the evaluation artifact. For example, district training for evaluators in time management suggests awareness by the district of school-level reactions to the time-consuming nature of the evaluation process. In addition, the politics of supervisor-teacher relations at the school level shaped the nature of written evaluation feedback, which was almost uniformly positive, even when teachers received relatively low scores on specific rubrics.

Recognition of the political nature of evaluation might help to untangle how issues of training and skill development combine with existing political and cultural expectations for the evaluation process. Political response is rational and appropriate if it facilitates implementation of the evaluation system. Recognition of the political nature of implementation could enable district and school leaders to view political response as a part of the process on the way to full implementation of the evaluation artifact, and not the final destination. Explicit attention to the political nature of evaluation and an examination of the features of the evaluation artifact could enhance the ability to use evaluation to provide constructive feedback in a dynamic political and cultural organizational context.

The Valle Verde Unified approach to implementing a new teacher evaluation system relied on a low-stakes, developmental model that depended heavily on the ability of local evaluators to extend their prior evaluation experience to meet the requirements of the new

system. Our study of the resulting implementation suggests that the district has developed local capacity to use the framework for teaching to support richer teacher and leader interaction around instruction. While much sensemaking research looks backward to investigate the relation of the past to the present, our perspective suggests that a sensemaking perspective can also point to areas for subsequent development. Future research is needed to understand how leaders might choose and exploit the potentially transformative features of evaluation system and integrate these features into new practices of teacher evaluation.

Notes

[1] A previous version of this paper was presented at the 2003 Annual Meeting of the American Educational Research Association, Chicago, Illinois, April 2003. The research reported in this paper was supported in part by a grant from the U.S. Department of Education, Office of Educational Research and Improvement, National institute on Educational governance, Finance, Policy-Making and Management, to the Consortium for Policy Research in Education (CPRE) and the Wisconsin Center for Education Research, School of Education, University of Wisconsin-Madison (Grant No. OERI-R3086A60003). The opinions expressed are those of the authors and do not necessarily reflect the view of the National Institute on Educational Governance, Finance, Policy-Making and Management, office of Educational Research and Improvement, U.S. Department of Education, the institutional partners of CPRE, or the Wisconsin Center for Education Research.

[2] The authors would like to thank Gary Zehrbach, Bill Thornton, and Terry Fowler for their contributions to data collection and analysis on the project.

[3] The will, skill, and structure elements are adapted from Rowan (1996), who describes teacher knowledge and skills (skill), teacher motivation (will), and the situation or context in which teachers work (structure) as critical factors influencing teacher and student performance.

[4] Names of the school district, schools, and educators have been disguised.

References

Berends, M., Bodilly, S., & Kirby, S. N. (2002). Looking back over a decade of whole-school reform: The experience of New American Schools. *Phi Delta Kappan*, *84*(2), 168-175.

Bronfenbrenner, U. (1979). *The ecology of human development: Experiments by nature and design*. Cambridge, MA: Harvard University Press.

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press

Chinn, C. A., & Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research*, *63*, 1-49.

Cohen, D. K., & Barnes, C. A. (1993). Pedagogy and policy. In D. K. Cohen & M. W. McLaughlin & J. E. Talbert (Eds.), *Teaching for understanding: Challenges for policy and practice* (pp. 207-239). San Francisco: Jossey-Bass.

Confrey J. (1990). A review of the research on students conceptions in mathematics, science, and programming. In: Courtney C. (ed.) *Review of research in education* (pp. 3-56). American Educational Research Association.

Danielson, C. (1996). *Enhancing professional practice: A framework for teaching*. Alexandria, VA: Association for Supervision and Curriculum Development.

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

Darling-Hammond, L., & Ball, D. L. (1997). *Teaching for high standards: What policymakers need to know and be able to do.* [Electronic version]. National

Education Goals Panel, June, 1997. Retrieved September 12, 2003 from http://www.negp.gov/reports/highstds.htm

Darling-Hammond, L., Wise, A. E., & Klein, S. P. (1999). *A license to teach: Raising standards for teaching*. San Francisco, CA: Jossey-Bass.

Darling-Hammond, L., Wise, A. E., & Pease, S. R. (1983). Teacher evaluation in the organizational context: A review of the literature. *Review of educational Research, 53(*3), 285-328.

Davis, D. R., Pool, J. E., & Mits-Cash, M. (2000). Issues in implementing a new teacher assessment system in a large urban school district: Results of a qualitative field study. *Journal of Personnel Evaluation in Education*, *14*(4), 285-306.

Desimone, L (2002). How can comprehensive school reform models be successfully implemented? *Review of Educational Research*, *72*(3), 433–479

Elmore, R. (2002). Bridging the gap between standards and achievement: The imperative for professional development in education. Washington, DC: The Albert Shanker Institute.

Fischoff, B. (1975). Hindsight foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance*,*1*, 288-299.

Gallagher, H. A. (2002). Vaughn Elementary's innovative teacher evaluation system: Are teacher evaluation scores related to growth in student achievement? Madison: University of Wisconsin, Wisconsin Center for Education Research, Consortium for Policy Research in Education.

Gentner, D., Rattermann M. J., & Forbus, K. D. (1993). The roles of similarity in transfer: Separating retrievability from inferential soundness. *Cognitive Psychology, 25*(4), 524-575.

Gentner, D. & Stevens, A. L. (Eds.). (1983). *Mental models*. Hillsdale, NJ: Lawrence Erlbaum.

Gibson, J. J. (1986). *The ecological approach to visual perception*. Mahwah, NJ: Erlbaum.

Greeno, J. G. (1998). The situativity of knowing, learning, and research. *American Psychologist*, *53*(1), 5, 26.

Greeno, J. G., Collins, A. M., & Resnick, L. B. (1996). Cognition and learning. In D. Berliner & R. Calfee (Eds.), *Handbook of educational psychology* (pp. 15-46). New York: Simon & Schuster Macmillan

Hallinger, P., & Heck, R. H. (1996). Reassessing the principal's role in school effectiveness: A review of empirical research, 1980-1995. *Educational Administration Quarterly, 32*(1), 5-44.

Halverson, R. (2002). *Representing phronesis: Supporting instructional leadership practice in schools*. Evanston, IL, Northwestern University. Unpublished dissertation

Halverson, R., & Zoltners, J. (2001). Distribution across artifacts: How designed artifacts illustrate school leadership practice. Paper for the 2001 American Educational Research Association, Seattle, WA.

Hammer, D., & Elby, A. (2002). On the form of a personal epistemology. In B. K. Hofer, & P. R. Pintrich (Eds.), *Personal epistemology: The psychology of beliefs about knowledge and knowing* (pp. 169-190). Mahwah, NJ: Erlbaum.

Haney, W., Madaus, G., & Kreitzer, A. (1987). Charms talismanic: Testing teachers for the improvement of American education." In E. Z. Rothkopf (Ed.), *Review of Research in Education* (pp. 169-238). Washington, DC: American Educational Research Association.

Hazi, H. M. (1994). The teacher evaluation-supervision dilemma: A case of entanglements and irreconcilable differences. *Journal of Curriculum & Supervision, 9*(2), 195-216.

Ilgen, D. R., & Davis, C. A. (2000). Bearing bad news: Reactions to negative performance feedback. *Applied Psychology: An International Review*, *49*(3). 550-565.

Keisler, S., & Sproull, L. (1982). Managerial response to changing. *Administrative Science Quarterly*, *27*, 548-570.

Kimball, S.M. (2003). 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, 16(4), 241-269*.

Kozlowski, S.W.J., Chao, G.T., & Morrison, R.F. (1999). Games Raters Play: Politics, Strategies, and Impression Management in Performance Appraisal. In James W. Smither, (Ed.) *Performance Appraisal: State of the Art in Practice* (pp. 163-205). San Francisco, CA: Jossey-Bass.

Longenecker, C. O., Sims, H. P., & Gioia, D. A. (1987). Behind the mask: The politics of employee appraisal. *The Academy of Management Executive*, *1*(3), 183-193.

Loup, K. S., Garland, J. S., Ellett, C. D., & Rugutt, J. K. (1996). Ten years later: Findings from a replication of a study of teacher evaluation practices in our 100 largest school districts." *Journal of Personnel Evaluation in Education*, *10*, 203-226.

March, J. G., & Simon, H. A. (1958). *Organizations*. New York: Wiley.

Milanowski, A. T., & Heneman, H. G., III. (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.

Murphy, J. (1994). Transformational change and the evolving role of the principal: Early empirical evidence. In J. Murphy & K. S. Louis (Eds*.). Reshaping the principalship: Insights from transformational change efforts* (pp. 20-53). Newbury Park, CA: Corwin.

Murphy, K. R., & Cleveland, J. N., (1995). *Understanding performance appraisal: Social, organizational and goal-based perspectives*. London: Sage.

Norman, D. A. (1993). *Things that make us smart: defending human attributes in the age of the machine*. Reading, MA: Addison-Wesley.

Odden, A. and Kelley, C. (2002). *Paying teachers for what they know and do: New and smarter compensation strategies to improve schools*. Thousand Oaks, CA: Corwin Press.

Pea, R. D. (1993). Practices of distributed intelligence and designs for education. In Salomon, G. (Ed.). *Distributed cognitions: Psychological and educational considerations* (pp. 47-87). Cambridge, UK: Cambridge University Press.

Peterson, K. D. (1989). Secondary principals and instructional leadership: Complexities in a diverse role. Madison, WI: The National Center for Effective Secondary Schools.

Peterson, K. D. (1995). *Teacher evaluation: A comprehensive guide to new directions and practices*. Thousand Oaks, CA: Corwin.

Ross, B. H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 629-639.

Rowan, B. (1990). Commitment and control: Alternative strategies for the organizational design of schools. Review of research in education. American Educational Research Journal. *16*, 353-389.

Rowan, B. (1996). Standards and incentives for instructional reform. In S. H. Fuhrman & J. O'Day (Eds.), *Rewards and reform: Creating educational incentives that work* (pp. 195-225). San Francisco: Jossey-Bass.

Schank, R. C. (1982). *Dynamic memory: A theory of reminding and learning in computers and people*. New York: Cambridge University Press.

Schank, R. C., & Abelson, R. P. (1977). Scripts, plans, goals, and understanding. Hillsdale, NJ: Erlbaum.

Simon, H. A. (1996). *The sciences of the artificial*. Cambridge, Mass., MIT Press.

Spillane, J., Halverson, R., Diamond, J. (2001). Investigating school leadership practice: A distributed perspective. *Educational Researcher*, 30(3), 23-28.

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

Stake, R. E. (1995). *The art of case research*. London: Sage.

Starbuck, W. & Milliken, F. (1988). Executives' perceptual filters: What they notice and how they make sense. In Hambrick, D (ed.) *The Executive effect: Concepts and methods for studying top managers* (pp. 35-65). Greenwich, CT: JAI.

Wartofsky, M. W. (1979). *Models: Representation and the scientific understanding*. Dordrecht, Holland ; Boston, D. Reidel.

Weick, K. E. (1976). Educational organizations as loosely coupled systems. *Administrative Science Quarterly, 21*(1): 1-19.

Weick, K. E. (1996). *Sensemaking in organizations*. London, Sage.

Wise, A. E., Darling-Hammond, L., McLaughlin, M. W., & Bernstein, H. T. (1984). *Teacher evaluation: A study of effective practices*. Santa Monica, CA: Rand.

Wright, P.S., Horn, S.P., & Sanders, W.L. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education, 11,* 57-67.

Appendix A

District Overview and Evaluation System Summary

The school district is the second largest in the state and includes 85 schools, approximately 60,000 students, 3,700 certified staff, and 270 administrators. Thirty-eight percent of the student population is non-white, with Hispanic students making up the largest part of the non-majority group. Although the district had recently revised aspects of its teacher evaluation system, the district and teachers' association agreed in 1997 that more comprehensive evaluation reforms were needed.

The new teacher evaluation system includes all of the standards and many of the suggested sources of evidence included in the *Framework for Teaching* (Danielson, 1996). There are four domains of practice with 23 components and 68 elements elaborating behavioral descriptions of the components. The domains are Planning and Preparation, Classroom Environment, Instruction, and Professional Responsibilities. Each element includes separate descriptions of teaching performance on a four-level rubric: unsatisfactory, target for growth (level 1), proficient (level 2), and area of strength (level 3). Table 1 includes an example of one set of rubrics for one of the 68 elements.

Table 1
Example of Rubric for Domain 1: Planning and Preparation; Component 1b: Demonstrating Knowledge of Students

Element	Unsatisfactory	Target for Growth/Level 1	Proficient/Level 2	Area of Strength/ Level 3
Knowledge of Students' Varied Approaches to Learning	Teacher is unfamiliar with the different approaches to learning that students exhibit, such as learning styles, modalities, and different "intelligences."	Teacher displays general understanding of the different approaches to learning that students exhibit, and includes a limited variety in lesson planning.	Teacher displays solid understanding of the different approaches to learning that different students exhibit and occasionally uses those approaches.	Teacher uses, where appropriate, knowledge of students' varied approaches to learning in instructional planning, as an integral part of their instructional planning repertoire.

Multiple sources of evidence are called for to assess performance relative to the standards. Evidence may include a teacher self-assessment, a pre-observation data sheet (lesson plan), classroom observations, pre- and post-observation conferences, other observations of teaching practice (e.g., parent-teacher meetings or collegial discussions), samples of teaching work and instructional artifacts, reflection sheets, three-week unit plan, and logs of professional activities. Unlike the suggestions in the Framework for Teaching, instructional portfolios are not required as part of the evaluation evidence.

Similar to the district's prior system, teachers are evaluated annually and specific procedures exist depending on where teachers are in three evaluation stages: probationary, post-probationary major, and post-probationary minor. Probationary teachers are those who are novice teachers or who taught previously in another district. Probationary teachers are observed at least nine times over three periods of the year and are provided a written evaluation at the end of each period. If they don't advance after their first year, probationary teachers undergo a second probationary year. If there performance is unsatisfactory, they may be dismissed.

Teachers in post-probationary status are evaluated in a major evaluation based on two of the performance domains, one selected by the teacher and the other by the evaluator. Formal observations occur three times over the course of the year and a written evaluation is provided toward the end of the year. After successful major evaluation, teachers move to a two-year minor evaluation phase.

Teachers on the post-probationary minor cycle are evaluated on one domain and receive one formal observation, resulting in one written evaluation at the end of the year. The process is repeated during the next year, with one new evaluation domain selected. An

optional minor evaluation process is available to teachers who have at least five years experience in the district and have been successfully evaluated under the major phase. Teachers in the alternative minor process may choose from six professional growth options (e.g., pursuit of National Board Certification, supervising a student teacher or engaging in an action research project). These options must still be tied to an evaluation domain, but are less structured than typical minor evaluations.

The written evaluations include a cover sheet with the teacher's name and basic demographic information (hire date, school, grade/subject, type of contract), whether the teacher is on the probationary and post-probationary cycle, and when the evaluation and observations occurred. Pursuant to state law, the form also indicates whether the complete evaluation was satisfactory or unsatisfactory. The form ends with evaluator and teacher signatures.

Evaluators are to mark the appropriate performance level on the four-level rubric for each element of each domain evaluated. Following the scores, the evaluators are required to provide a narrative description of the evaluation. The narrative is to include a separate description for any element receiving an unsatisfactory rating, with evidence cited from observations, and recommendations. For domains with scores above unsatisfactory (level 1-3), the form calls for "one complete narrative mentioning data for each, commendations, and recommendations." Any evaluation standard rated unsatisfactory results in an unsatisfactory evaluation and the teacher undergoes an intervention process. Teachers in the intervention process work with their administrator to establish an assistance plan and are evaluated on all performance standards that are not being satisfactorily met. When all objectives of the assistance plan are met, teachers may go back into the regular evaluation cycle.