Educational Administration Quarterly
Vol. 39, No. X (Month 2006) 1-

Evaluation in the Wild: A Distributed Cognition Perspective on Teacher Assessment

Richard R. Halverson
Matthew A. Clifford

Purpose: *The authors adapted distributed cognition theory to provide a detailed account of how school leaders use knowledge of the new programs, existing initiatives, and the school context to guide policy implementation in local school contexts.*

Research Design: *The study used distributed cognition theory to show how policy implementation studies provide an occasion to understand the influence of context on practice. The article focuses on a case study of (a) a suburban district design of a teacher evaluation policy and (b) a principal's effort to use the evaluation program with the teachers in her middle school. The authors adapted the distributed cognition theory to provide an analytic framework to better address the issues of school leadership.*

Findings: *The authors found that the design of the policy required evaluators to address the tensions between summative and formative evaluation implicit in the program design. In this case, the principal relied heavily on her discretion to determine which features of the teacher evaluation policy would be emphasized with different teachers. The case also provided insight into how the principal reconciled the demands of evaluation with ongoing instructional and personnel demands.*

Conclusions: *The distributed cognition framework provides a valuable tool for organizing close studies of the cognitive and contextual dimensions of leadership practice and can provide valuable information about how policies can be designed and used to shape real changes in everyday practice.*

Keywords: *distributed cognition; instructional leadership; policy design and implementation; teacher evaluation; case study; empirical paper*

In his seminal work on distributed cognition, Edwin Hutchins (1995a) remarked, "Many of the foundational problems in cognitive science are consequences of our ignorance of the nature of cognition in the wild" (p. 370). Distributed cognition theory provides a set of promising conceptual and analytical tools for understanding the interaction of cognition and context in the

DOI: 10.1177/0013161X05285986

wilds of everyday practice. Hutchins's early research used distributed cognition theory to understand how pilots navigate ships and planes. Hutchins proposed the idea of a cognitive system composed of actors, tools, and designed contexts as the irreducible unit of analysis to understand cognitive tasks. Distributed cognition research proceeds by close study of how the micro tasks of a practice are completed both to learn how actors work and to determine guidelines for how to better design effective cognitive systems.

This article represents our initial effort to adapt the distributed cognition theory to study the central problems of school leadership practice. Our argument explores what a distributed cognition analysis can tell us about a chronic problem of leadership and teaching practice in schools: teacher evaluation. Despite the promise of distributed cognition theory, our research points to problems with its application to issues of leadership. The distributed cognition framework is still developing as a theory to study cognition in the wild. Early work in distributed leadership (e.g., Spillane, Halverson, & Diamond, 2001, 2004) relied on distributed cognition to frame an approach to understand school leadership. Still, distributed cognition theory has not yet been widely applied directly to the study of issues such as school leadership. In part, this is because the core concepts of distributed cognition have not been adequately adapted into a framework to guide organizational analysis. The first part of our article addresses this need by developing an analytic framework, based on a literature review of distributed cognition, sense making, and policy implementation, appropriate for studying issues of school leadership. Our distributed cognition framework provides a principled method for investigating how actors construct tasks in terms of their perceived capacity of the existing cognitive system.

In the second part of the article, we apply the framework to tackle a chronic leadership issue in schools: teacher evaluation. Although teacher evaluation has great promise for improving teaching in learning in schools, evaluation practices have notoriously been thwarted by the organizational and professional context of schools. The distributed cognition framework helps us access just how evaluation tools are used in practices. The case that we present details how a school leader made sense of reform-based teacher evaluation practices in terms of the school's existing cognitive system. As a result of our analysis, we show how a leader moved beyond the central aims of the evaluation policy and used the policy tools in terms of her perceived needs for the school. We demonstrate how discretion emerged as a central cognitive activity in teacher evaluation and argue that the ability to repurpose cognitive artifacts on the fly to satisfy multiple organizational needs is a significant form of leadership expertise. Although some educators and reformers may conclude that our analysis describes a case of stunted reform, we feel

that the distributed cognition framework reveals just how contexts contribute to change and points toward how policy designers can attend to the conditions of existing cognitive systems to effect lasting improvements in schools.

POLICY IMPLEMENTATION AND TEACHER EVALUATION

In schools, as in other organizations, policies are drafted to influence the practice of others. The road from intention to outcome, however, is rarely straightforward. Early policy implementation studies showed that local situations shape how policies are used (Lindblom, 1995; Lipsky, 1980). Policies drafted to change institutional practices rely on the active participation of local actors who highlight, redesign, and transform certain policy features into practice. Research in this tradition has also recently emphasized the cognitive aspects of implementation, arguing that the cognitive frames and affective expectations of local practitioners influence which policy features are emphasized and which are ignored (Spillane, Reiser, & Reimer, 2002; Starbuck & Milliken, 1988). Sense making dominates the thinking of practitioners because previously implemented policies and programs combine with institutional traditions to establish rich, and stubborn, networks of interconnected practices (Talbert & McLaughlin, 1993). Sometimes this process of mutual adaptation (McLaughlin, 1987) results in local practitioners' capturing the essence of the policy; other times, implementation results in a "lethal mutation" (Brown & Campione, 1996) that may reflect surface features but omit the essential, underlying heart of the policy. Local school actors judge the value of new policy features against the perceived requirements of these aggregated policies and practices (Halverson & Clifford, 2004; Spillane & Thompson, 1997). New policies that require more resources than local practitioners see as available are often "satisficed" (Simon, 1997) in terms of existing constraints.

Reform-based teacher evaluation artifacts provide a unique opportunity to examine how the hopes of policy design meet the realities of existing practice. On one hand, teacher evaluation programs promise the ability to formatively and summatively assess new practices in terms of desired outcomes. Clear, legitimate access to teaching is necessary for supporting teachers to improve practice. Evaluation programs also provide accountability measures necessary to address staff quality issues and to provide grounds for dismissing poor teachers. In practice, however, the summative and formative functions of assessment are often set against each other to undermine the potential effects of evaluation (Natriello, Pallas, & McDill, 1990). The traditional

opposition of administration and teaching practice severely curtails the potential of teacher evaluation (see, e.g., Hazi, 1994; Sergiovanni & Starratt, 1993). When teacher evaluation is aimed at summative quality control, formative practices often drop out and teachers end up isolated in classrooms with little valuable feedback. Teacher assessment is then used to "weed out" poor performing teachers rather than to hold all teachers accountable or to improve the performance of all teachers (Darling-Hammond, Wise, & Klein, 1999; Haney, Madaus, & Kreitzer, 1987). Furthermore, most current teacher evaluation practices fail to provide sufficient training and lack the support of teachers and school leaders (Loup, Garland, Ellett, & Rugutt, 1996; Peterson, 1995). The resulting watered-down, marginalized teacher assessment practices are typically the product of a complex network of trade-offs, as practitioners adjust policies to the constraints of the existing situation. This tension between formative and summative pressures continues as teacher evaluation practices currently transition to shared leadership models that involve coaching and mentoring (Mangin, 2004). Even practitioners perceived as successful implementers of standards-based teacher evaluation practices need to engage in trade-offs as they adjust the demands of the new policy artifacts to the needs of their existing contexts (Halverson, Kelley, & Kimball, 2004; Kimball, 2003; Milanowski & Heneman, 2001). The tendency of teacher evaluation practices to run headlong into the traditions of local practice provides a prime opportunity to study how practitioners make sense of the new in terms of the old.

A key research challenge for understanding how teacher evaluation policies influence practice is to access how evaluators adapt evaluation tools to the needs of particular teachers. This need is particularly acute for principals who must balance summative and formative feedback within the same evaluation cycles. Traditional approaches to teacher evaluation research at the level of policy, document, or outcome analysis do not provide access to appropriate levels of practice where teacher evaluators make their rating and feedback decisions. We need tools to help us understand how and why evaluators make sense of evaluation tools in the black box of practice then connect the micro study of practice back to the policy picture.

DISTRIBUTED COGNITION

In recent years, learning scientists have developed several new frameworks to study how thinking and learning occur in complex environments (Cole, Engeström, & Vasquez, 1997; Hutchins, 1995a, 1995b; Lave & Wenger, 1991; Rogoff, 1990; Salomon & Perkins, 1993; Wertsch, 1998).

One framework, distributed cognition, was designed to trace the influence and interconnection of tools and thinkers in complex learning environments (Hutchins, 1995a; Pea, 1993; Perkins, 1993). Distributed cognition theory grew from research in human-computer interaction (Halverson, 1995; Hollan, Hutchins, & Kirsch, 2000; Zhang & Norman, 1994) and studies of professional practices (Dunbar, 1995; Goodwin, 1995; Lave, 1988; Neressian, Kurz-Milcke, Newstetter, & Davies, 2003). Distributed cognition theorists followed the lead of Leon'tev (1975, 1981) and Vygotsky (1978) to turn the existing model of cognitive analysis inside out: Instead of dwelling on cognition in the head, distributed cognition theorists focused on operation of cognition in the world. Hutchins (1995a) explained,

> Thinking about organizations as cognitive systems is not new . . . what is new is the examination of the role of the material media in which representations are embodied, and in the physical processes that propagate representations across media. Applying the cognitive science approach to a larger unit of analysis requires attention to the details of these processes as they are enacted in the activities of real persons interacting with real material media. (p. 266)

Hutchins observed that taking cognition in the world as the unit of analysis allows researchers to attend to aspects of cognition that can be inferred only when the unit of analysis is the individual. If intelligence is better seen as an achievement rather than as a possession, as suggested by Roy Pea (1993), then studying the systems that support achievement offers new insight into the cognition of actors in organizations.

Hutchins's work shows how a distributed cognition perspective reveals cognition in context. In sociotechnical systems, artifacts provide manageable representations of complex data that reduce human cognitive loads and aid communication. In "How a Cockpit Remembers Its Speed" Hutchins (1995b) analyzed the task of piloting a passenger jet to show how speed control reveals the distribution of cognition among pilots and cockpit instruments. Hutchins found that the organization of artifacts in the work environment distributes cognition for actors both temporally (creating memory structures that offload cognitive demands in high-cognitive load activities) and socially (allowing actors to communicate understandings through shared representations). A frequent finding in a distributed cognition analysis is that seemingly innocuous artifacts often play critical and unacknowledged roles in supporting key task. Hutchins, for example, showed how "speed bugs," the interactive devices attached to the rims of analog speed and altitude gauges, allow pilots to easily access fuel and plane capability information contained otherwise on separate, difficult-to-access index cards. Speed bugs represent

the relevant information ready to hand, distributing the cognitive labor so that pilots can concentrate on landing the plane rather than looking up number tables. Tracing a task through a complex system reveals which artifacts structure tasks and articulates the tacit assumptions actors make operating within the cognitive system. Without reference to the collection of artifacts (gauges, bugs, cards, and controls) in the cockpit, we would miss the cognitive activity that guides the task of flying.

Here we push Hutchins's initial conception of distributed cognition to build an analytic framework appropriate for studying the policy and practice of teacher evaluation. In adapting the distributed framework for analyzing policy development and implementation, we follow Cohen and Hill's (2001) suggestion to distinguish the policy from the instruments (artifacts) deployed in its support. This enables us to analyze the artifacts provided by the policy and to address the range of artifacts used by practitioners in implementing a policy. Three key questions guide our distributed cognition analysis: (a) What is the task? (b) What are the relevant artifacts? (c) How are tasks and artifacts linked in a cognitive system? We discuss how each question flows from distributed cognition theory below, and then in the case that follows, we use these questions to organize our distributed cognition analysis of teacher evaluation practice.

What Is the Task?

A distributed cognition perspective focuses on how tasks flow through complex systems. A task is a basic building block of practice, a discernable sequence of behaviors that helps actors accomplish goals. Tasks can be described at different grain sizes: macro tasks involve descriptions at the large-scale organizational level, whereas micro tasks describe the specific behaviors involved in executing macro tasks (Spillane et al., 2001). Hutchins's analysis in *Cognition in the Wild* (1995a), for example, focuses on the macro task of navigation to identify the micro tasks that set the ship's course. From a school leadership perspective, macro tasks such as "monitoring of instruction" or "establishing a vision" are composed of micro tasks such as "talk to Ms. Freeney in the guidance office this morning about the attendance reports." The choice of task to study guides the features of the cognitive system to be uncovered.

What Are the Relevant Artifacts?

Focusing on how key tasks are enacted reveals the relevant structural supports, or artifacts, that support practice. The origin of the term *artifact*

reaches back to Aristotle's insight that the product of an artisan's work is composed of matter formed by the intent of the designer (Aristotle, 1941). Artifacts can range from tangible objects such as hammers, pans, or calculators to abstract entities such as policies, programs, or procedures. Distributed cognition analyses use the concept of *cognitive artifacts* (Norman, 1991) to show how artifacts carry meaning and support communication. The artifact design includes features that reflect the intentions of the designer on suggested uses or effects. Roy Pea (1993) noted how artifacts "represent some individual or community's decision that the means should be reified as a quasi-permanent form for others" (p. 53). Because cognitive artifacts are carriers of previous reasoning, artifact use represents a kind of asynchronous communication between the designer and the user. In other words, artifact use can be understood as a form of human interaction (Pea, 1993).

We use the term *artifact* instead of more commonly used terms such as *policy* or *program* to provide a general descriptive term for the range of tools leaders use to shape practices in schools. Schools rely on a wide range of inherited and locally designed artifacts such as daily schedules, budgets, curricula, and report cards to organize the work of teaching and learning (Halverson, 2002, 2004). The work of policy makers in education can be seen as inscribing intentions into policy artifacts through designed features with the hope that practitioners pick up on these features to shape practice. Leaders design and deploy networks of artifacts to influence the practice of others (Halverson, 2003; Spillane et al., 2004).

Artifacts provide a path to highlight cognitive aspects of how actors interact with structures. Artifacts are usually designed to contain certain features, such as specific instructions, rewards, and shortcuts that facilitate their use. Simply listing features does not help us understand how artifacts are used. Task analysis helps identify which artifact features actors select and how features adapted to achieve ends. Hutchins's (1995a) analysis of ship navigation, for example, considers how artifacts such as written procedures are developed by designers to specify tasks and assign responsibilities (p. 295). Once the macro tasks (navigation) are analyzed to identify the main artifacts (policies, instruments, and maps), the process proceeds to analyze which micro tasks will be able to reveal how artifact features are actually used in practice.

Sense making plays a key role in artifact use. Artifact features are often interpreted differently than the designers might have intended. Actors make sense of new artifacts by selecting appropriate artifact features according to the social context of use, the experience of the user, and the perceived needs of a specific occasion (Spillane et al., 2002). For example, artifacts that influence work practices are often received into (or generated from) communities

of practice (Lave & Wenger, 1991) that help practitioners judge which features are relevant and actionable (Halverson & Zoltners, 2001). Although Pea (1993) may have overstated the situation with his comment that "inscriptions rarely reveal their affordances for activity" (p. 62), artifact designers have long recognized the tenuous connection between artifact affordances and resultant action. The difficulty of tracing the effects of artifacts on practice rests in part on our inability to anticipate the cognitive contribution of actors. A distributed cognition analysis helps determine this cognitive aspect of implementation by attending to which artifacts matter and on how contexts matter to practitioners' artifact feature selection.

How Are Tasks and Artifacts Linked in a Cognitive System?

The cognitive system refers to the aggregated networks of artifacts and actors necessary to complete complex organizational tasks. Tasks are mediated by networks of artifacts that, in turn, establish the range of practices in an organization. The cognitive system provides a unit of analysis with several interesting properties. First, focusing on the cognitive system takes the focus off the actor's cognition. Traditional leadership studies that focus on the knowledge and skills of individual actors miss the critical interactive aspects of how actors and artifacts together constitute practice. As Hutchins (1995a) noted,

> If we ascribe to individual minds in isolation the properties of systems that are actually composed of individuals manipulating systems of cultural artifacts, then we have attributed to individual minds a process that they do not necessarily have, and we have failed to ask about the processes they actually must have in order to manipulate these artifacts. This sort of attribution is a serious but frequently committed error. (p. 173)

Paradoxically, a second advantage of focusing on the cognitive system is to shed new light on practitioner cognition. Instead of attempting to peer inside the actor's head and attribute intentionality to behavior, a distributed cognition analysis sees the actors' actions as the thinking of the cognitive system. Cognitive systems, whether inside or outside the head, include input mechanisms, memory structures, rules for decision making, and output mechanisms. In individuals, the existence and operation of these components must be inferred from inaccessibly nested neurological and psychological structures. From a distributed cognition perspective, however, the memory

structures and rule systems of the cognitive system are externalized, and accessible, through artifacts. Artifact feature selection is, from the perspective of the cognitive system, an analogous process to what goes on in the head. But feature selection and adaptation are visible manifestations of cognitive activity closed to traditional cognitive analyses. Hutchins's research shows how the thinking of the cognitive system is seen, for example, through setting speed bugs to serve as memory devices for landing a plane. In other words, the thinking of a cognitive system is displayed though how actors interact with artifacts to complete tasks. The connections between tasks, artifacts, actors, and the cognitive system serve to connect the analyses of micro tasks to the macro tasks of practice, providing valuable insights for practitioners as well as policy designers.

METHOD

Case studies have proven particularly useful for examining how multiple variables interact within an environment (Merriam, 1998). Our case study shows how a middle school principal conducted her teacher evaluations over the course of a year with a new district-designed, standards-based teacher evaluation artifact. The Stillwater district (a pseudonym) serves 2,900 students in four elementary schools, two middle schools, and one high school. We chose Stillwater because of its recent efforts to develop an innovative teacher evaluation program and its willingness to participate in the research project. Although we decided to focus the case presented here on the practice of one of our participants, a middle school principal, we also include information from our work with the elementary school principal and interviews with district leaders.

We negotiated for several months with district leaders to provide access to observe the evaluation system in action in a middle school and an elementary school. In the end, we collected several types of data.

1. We interviewed members of the district teacher evaluation design team, including the superintendent, the director of curriculum, and several principals, about the design and the implementation of the teacher evaluation system.
2. We followed the elementary and middle school principals through all or part of their evaluation practices for 16 teachers (8 in the middle school, 8 in the elementary school). (Because each principal was responsible for 20 teacher evaluations per year, we captured all or part of 40% of the teacher evaluation work for each principal during the 2002-2003 school year.)

In tracing the evaluation process, we shadowed the principals during the actual classroom observation, videotaped the principal-teacher postobservation conferences, and interviewed teachers and principals after the conference. In the end, we collected 11 complete cases of teacher evaluation practice from preobservation goal setting though the postobservation conferences and interviews.

The case we present here results from a qualitative data analysis of video, text, and interview transcripts facilitated by Atlas.ti software. We used the distributed cognition framework to code for the relevant tasks, artifacts, and features of the cognitive system. We identified the kinds of interactions teacher and principals mediated by the artifacts and linked the resulting micro tasks into depictions of the evaluation macro task. We also analyzed the conference conversations to measure duration and frequency of interactions. We then shared our preliminary case drafts with evaluators to test the quality of our representations and to correct our errors and oversights.

For the purpose of this article, we chose to focus on the middle school principal, Karen Page, to provide an in-depth analysis of how a specific leader engaged in evaluation tasks. Although both principals provided rich data about their evaluation practices, we chose Page because her school allowed us access to a wider range of teachers and staff evaluation practices (including guidance counselors and special education staff as well as classroom teachers) and we felt that the instructional organization of the middle school, which involved both specialization and grade level cross-disciplinary coverage, could contribute to our understanding of how teaching and learning might be addressed across different levels of schooling. Our elementary principal analysis was used to supplement our case and to provide a sense of context for the middle school principal's work using the district's evaluation practices.

The case design addresses the three main questions of the distributed cognition framework: (a) What is the task? (b) What are the relevant artifacts? and (c) What is the cognitive system? Much of the case focuses on unpacking the second question. After identifying the key artifact as the district teacher evaluation policy, we analyze how different stakeholder perspectives in the design process influenced the final artifact design. We then present an analysis of the micro tasks of evaluation to identify the range of artifact brought into play during the evaluation process. The case concludes with a consideration of a comprehensive school reform model, the central artifact of the Baxter cognitive system that influenced Page's teacher evaluation practice.

EVALUATION IN THE WILD: THE CASE OF BAXTER MIDDLE SCHOOL

This case follows middle school principal Karen Page through her 2002-2003 teacher evaluation practice. We found that the implementation of the teacher evaluation artifact was far from a simple process. The evaluation task was composed of five distinct micro tasks, each supported by a separate set of artifacts. Some artifacts were provided by the main district policy artifact, the Teacher Professional Growth Program (TPGP), but the evaluator contributed others. We also found that a key artifact of the school's cognitive system, a comprehensive school reform plan called Expeditionary Learning (EL), played a significant role as a sense-making filter for the principal and the teachers. Although EL acted as a sense-making filter for the school community, we found little evidence of a designed, programmatic connection between EL and TPGP. The principal seemed to use the TPGP as a tool to support her perceived role as the school's instructional leader by balancing TPGP requirements with her perception of individual faculty member needs and the goals of the existing cognitive system.

The Baxter Middle School serves about 700 sixth- to eighth-grade students from the surrounding middle-class neighborhood. Principal Page served as the main evaluator in her school. A 28-year veteran educator, Page spent all but 2 years of her career at Baxter. Page was in her 2nd year as principal during our research after 2 years as assistant principal and 23 years in the classroom. She played a significant role on the district teacher evaluation design team and believed that teacher evaluation could be an important way of strengthening principal-teacher relationships necessary for instructional leadership. She viewed teacher evaluation as partially fulfilling her duty to her community by ensuring the highest quality teacher works in each Baxter classroom.

What Is the Task?

The teacher evaluation practices in the Stillwater district represented typical practice for many American schools. The evaluation process was structured to allow for summative teacher rating as well as formative discussions of teaching practice between principals and teachers. The features of the evaluation artifact were established in negotiations between the teachers union and district leaders. The district design of the evaluation system structured the macro task of evaluation into a number of distinct micro tasks

through a series of forms provided by the district. The evaluation tasks consisted of a preobservation session, a classroom observation, and postobservation discussion. Prior to the observation, the evaluator discussed lesson plans and points of emphasis with the teacher. The principal typically observed the teacher for a class period and then completed a district-mandated checklist of expected behaviors and a narrative of the observation. The teacher and principal then met to discuss the observation and resultant ratings. The sequence ended with the teacher signing off on the written evaluation. The written evaluation provided an evidentiary basis for determining professional advancement for the teacher. This process was repeated three times during the school year for novice teachers and several times for struggling tenured teachers.

Completing the teacher evaluation cycle represented a significant time commitment. The elementary and middle school principals, each responsible for evaluating 20 teachers, spent about 2 to 3 hours on each evaluation write up and another 2 to 2½ hours in postobservation meetings. Six teachers in each school received three observations for the year, whereas the remaining teachers were evaluated once. Probationary teachers invested 1 to 2 hours completing self-evaluation and goal-setting forms and 4 to 6 hours in multiple observation conferences over the course of the year; postprobationary teachers spent the same time with the forms but only 1 to 2 hours for postobservation conferences. These time estimates suggest that principals spent between 100 and 150 hours, or somewhere between 7% and 10% of their professional time, during the 2002-2003 school year engaged in the evaluation process.

What Are the Relevant Artifacts?

The next step in our analysis involves identifying the key artifacts involved in the task of evaluation. We found one central artifact provided by the district to guide the macro task of evaluation: the Stillwater Teacher Professional Growth Program. Our account of how the TPGP was used to guide teacher evaluation at Baxter begins with the story of the district artifact design process. In the sections that follow, we analyze the evaluation micro tasks to reveal the artifacts used to guide evaluation practice in the wild.

Macro Task Artifact: Stillwater TPGP

This TPGP artifact was the result of a district evaluation program redesign during 2000-2002. The district administrator related how, in the late 1990s, the Stillwater district faced public and school board pressure to revamp the

existing teacher evaluation system. Labor conflicts brought issues of teacher accountability to the front of the agenda and pressed for a comprehensive, standards-based evaluation system. In late 2001, the superintendent and the director of curriculum responded by calling together a design team of principals, teachers, and staff members to redesign the system. After visiting several other districts for design ideas, the design team selected Charlotte Danielson's (1996) Framework for Teaching as their model for the district to assess teachers across well-defined performance levels.

The Danielson framework is organized into four domains: Planning and Preparation, the Classroom Environment, Instruction, and Professional Responsibilities (see the appendix). Each domain is organized into several components; in turn, the components are broken into specific elements. The Instruction domain, for example, contains five components (such as Communicating Clearly and Accurately and Engaging Students in Learning) with three to four elements per component (for an example, see Table 1). Each element includes rubrics to specify unsatisfactory, basic, proficient, and distinguished performance.

The Stillwater design team met monthly for a year and a half (during the 2000-2002 school years) to adapt the Danielson framework into the TPGP. The collaborative design team employed a stakeholder strategy to encourage buy in from district leaders, principals, and teachers. Each group wanted to make their mark on the final design, and the underlying disagreements about the role of teacher evaluation resulted in design trade-offs. The superintendent, for example, perceived his role as resolving long-standing labor-management issues in the district to refocus district efforts on student learning:

> There was no respect at all, so . . . everybody walked in with that type of a mind-set, you know, that we were going to go to battle on this, and then that of course carried on to everything else. We worked hard at trying to change that mind-set and engaged in a collaborative bargaining process that resulted in a voluntary agreement for the first time in 20 years here.

He described how the TPGP needed to be recognized as an accepted measure of teacher quality that signaled the district's dedication to high-quality instruction to the community while also providing tools for teachers and leaders to build trust within schools. The teacher evaluation process had come up as a contentious issue because district and school board officials were concerned that veteran teachers were rarely evaluated. In the TPGP design, postprobationary teachers were to be regularly evaluated but less often than probationary teachers.

TABLE 1.
Stillwater SAR Instructional Domain—Domain 3: Instruction, Component 3b:
Using Questioning and Discussion Techniques, Elements: Quality of Questions, Discussion Techniques

	Level of Performance			
Element 1	*Unsatisfactory*	*Basic*	*Proficient*	*Distinguished*
Quality of questions	Teacher's questions are virtually all of poor quality	Teacher's questions are a combination of low and high quality; only some invite a response	Most of the teacher's questions are of high quality; adequate time is provided for students to respond.	Teacher's questions are of uniformly high quality with adequate time for students to respond; students formulate many questions
Discussion techniques	Interaction between teacher and students is predominantly recitation style with teacher mediating all questions and answers	Teacher makes some attempt to engage students in a true discussion with uneven results	Classroom interaction represents true discussion with teacher stepping, when appropriate, to the side	Students assume considerable responsibility for the success of the discussion, initiating topics, and making unsolicited contributions

District leaders also sought to strengthen the remediation aspect of the prior evaluation program to give greater latitude for dismissing poor teachers. The director of instruction felt that the Danielson framework improved on the prior system:

> [The old system had] a breakout of particular skills . . . [but with] no explanatory information or rubric to go with it. One of the elements would be . . . communicating with parents. Well what's a superior rating of communicating with parents look like? What does unsatisfactory look like? And you're just supposed to know. So what happens over time is over time you do get a sense of building your own personal rating system because over time you're working with teachers and all of a sudden you go wow, that's outstanding, that must be superior.

The district designers also discussed strengthening the TPGP incentive system by tying teacher bonuses to improve ratings or developing a system to correlate ratings with measures of student learning. These hopes, however, ran into significant opposition by teachers and leaders.

The teachers on the design team pushed for a stage model with different expectations for probationary, postprobationary, and struggling teachers. Teachers also argued that the standardization in the evaluation framework could restrict teacher and evaluator autonomy. The TPGP design addressed this perspective by including a teacher-directed goal-setting process for setting individual learning plans and a self-evaluation form for teachers to assess their own practice. Teachers also wondered about the objectivity of the TPGP, even given the inclusion of the Danielson rubrics. One teacher commented that "it's all very subjective, that's what I think about it, and I might fill it out differently on one day than the next. I don't stress over it, I just do it."

The principals involved in the design process had a different perspective. Several principals commented on the challenge of using the same process to provide both summative and formative feedback to teachers. One principal noted that

> The whole thing that stinks about evaluation is that you have to be the coach and the ref at the same time. I graded English papers for years, and I had the same complaint. You say "Try try this, do this," and in the end you slap it on them. They are incompatible roles in many ways.

Other principals challenged the district intention to develop a system for compiling and comparing teacher ratings, wondering how these collected ratings would be used and arguing that the existence of ratings might disrupt the sense of community in the schools. Principals fought to include a narra-

tive component in the TPGP to supplement the numerical rating to allow evaluators to explain their ratings and relate the teacher's value to the school. One principal commented that in her experience as a teacher, the main value of evaluation was to discuss practice with the evaluator:

> The most meaningful part was the conversation where [the principal] and I sat down and I heard that I was doing a good job, but the paperwork part didn't have much meaning. And I didn't expect anything more.

The evaluation instrument was ultimately designed to accommodate these multiple stakeholders' goals and perceived conflicts with current school reform, evaluation, human resource, and instructional systems. The TPGP addressed these different stakeholder perspectives by included artifacts that structure the evaluation process. The TPGP program was distributed as a binder organized into three main sections.

- Stage 1 described the evaluation process for probationary teachers,
- Stage 2 was for postprobationary teachers, and
- Stage 3 outlined the remediation and dismissal process.

Each stage consisted of a sequence of artifacts (mainly forms) to guide the evaluation micro tasks of goal setting and self-rating, preobservation planning, and the formal evaluation write up. The artifacts mainly guided teachers through the evaluation process or reported the evaluation results; no artifacts were included in the TPGP to specifically guide the practice of evaluators. These omissions required evaluators to rely on their experience with previous evaluations or artifacts from other evaluation programs to supplement the process specified by the TPGP, thus opening the door for artifact adaptation.

The foundational artifact of the TPGP was the Summative Assessment Report (SAR). The SAR contained two main components: a rating table based on the Danielson framework and a comment section for evaluators to provide an evaluative narrative. The rating tables for the self-rating and the SAR were adapted from the Danielson framework with one significant change: the elements did not include much space for evidence to justify ratings. Instead, each rubric dimension (from unsatisfactory to distinguished) was split into three sections, resulting in a 12-point checklist range. Although Danielson (1996) discussed the importance of providing adequate evidence to justify a rating within each element, the design of the Stillwater SAR provided a place for scores without room for including relevant evidence for each element. Although the comment section of the SAR was a space for

evaluators to provide a narrative explanation (and presumably to discuss appropriate evidence), there were no instructions provided for the content of the SAR comment section. Lacking space and direction, evaluators could use their discretion to determine what constituted credible evidence and include that evidence in the ratings and narrative.

Micro Task Analysis

We identified five micro tasks in our observation of teacher evaluation in the Stillwater district (see Table 2).

Micro Tasks 1 and 2: The Self-Evaluation and Preobservation Conference

The evaluation cycle opened with a teacher self-evaluation process and a preobservation conference. The TPGP provided two artifacts to structure the teacher's self-evaluation, a professional development plan and a self-evaluation form that uses the same format as the Danielson-inspired SAR. In the first stage of the evaluation process, Page distributed already-completed professional development plan forms to postprobationary teachers and then asked all teachers to rate themselves according to the self-evaluation form. Page then scheduled preobservation conversations and distributed the preobservation discussion forms several days prior to observations. The preobservation session also helped Page set her expectations for which evidence would be appropriate to include in the Instruction domain of the SAR.

Although the professional development plan, teacher self-evaluation form and preobservation discussion form served to structure the evaluation process for teachers, there were no corresponding artifacts provided to structure the initial process for evaluators. In the absence of artifacts to help focus the evaluator's attention in the preobservation conference, Page drew on her prior experience as a teacher and teacher evaluator to develop her own interview protocol for each preobservation conference, which included the following four questions: (a) What will I observe? (b) How did the lesson that I will observe arise? (c) How does this lesson fit into the overall curriculum? (d) Is there anything specifically that you want observed in the class?

The absence of artifacts to guide evaluators in the preobservation micro task was particularly relevant because of competing formative and summative functions of the TPGP artifacts. The professional development plan and the preobservation discussion form were designed as formative tools for teachers to determine the direction of their professional growth; the self-evaluation form was designed for teachers to summatively measure their

TABLE 2.
Teacher Professional Growth Program (TPGP) Tasks and Artifacts

Evaluation Task	*TPGP Artifact*	*Teacher Role*	*Evaluator Role*	*Time Requirement*	*Social Space*
Micro task 1 Self-reflection form preparation	Professional development plan; Self-assessment form; lesson plan form	Structured by the artifact	No artifact provided	About 1 hour for teacher to complete self-evaluation	Classroom
Micro task 2 Preobservation conference	Preobservation conference discussion form	Structured by the artifact	No artifact provided	30-minute teacher-principal meeting	Principal's office
Micro task 3 Classroom observation	No artifact provided	No artifact provided	No artifact provided	40-60 minutes	Classroom
Micro Task 4 Summative assessment report (SAR) preparation	SAR	None	Other than the SAR, no artifact provided to aid in calculating and preparing evaluation: calibrating assessment of instructional quality with principal assessment or for identifying ways to improve	1-2 hours to prepare SAR form	Principal's office; home
Micro task 5 Postevaluation conference	SAR	Self-evaluation structured by the artifact provided; however, sample ratings and evidence were not provided in the TPGP	Other than the completed SAR, no artifact provided to guide discussion	45 minutes for conference	Principal's office

performance in terms of established standards. Because these artifacts did not direct teachers about which elements to emphasize, teachers had the latitude to select goals according to their own, rather than the organizational, needs. The lack of direction designed into the artifacts placed the burden on evaluators to help teachers to link individual to systemic instructional goals.

Page's preobservation questions provided an occasion to explore teachers' thinking about curriculum and lesson design. These conversations allowed Page to draw on her knowledge of ongoing concerns for each teacher as a means to establish goals for the classroom observation. Page's questions pressed teachers to be explicit about their instructional goals. For example, discussion with one probationary teacher prompted Page to explore rubrics as an instructional technique.

> Page: How much experience have your students had in using rubrics?
> Teacher: They've used them two or three times, mostly in writing.
> Page: The reason that I ask . . . is that their prior experience using rubrics may have an impact on how quickly they get into this. That's not critical here because you don't have kids who haven't seen rubrics before this.

In a later interview, the same teacher commented he had not considered students' prior experience but that Page's comments sparked him to think about his approach to rubric design. With precious little time to formally meet with teachers, Page designed the preobservation sessions not only to plan for her observation but also to repurpose the time for catching up and to provide formative advice on other issues. Improvising with the scarce resource of conversational time allowed Page to interact with teachers while complying with the established purpose of the preobservation conference.

Micro Task 3: The Classroom Observation

The purpose of classroom observation is for the evaluator to gather the evidence for SAR ratings. The TPGP provided no artifacts to guide the evaluators in choosing evidence or making judgments during the observation. Although the district provided three in-services for evaluators to practice assessing a videotaped example of teaching practice, it did not provide any artifacts to carry these lessons over into classroom observations.

Page's typical observation practice included the following procedure: (a) With notebook in hand, select a seat toward the back of the classroom. (b) Sketch the room layout, noting student and teacher positions and room design features such as bulletin board placement and audiovisual equipment. (c) Outline the major lesson "moves" on the central portion of the page. (d)

Note questions about the teaching and student learning and comments on areas noted by the teacher (e.g., student engagement). (e) Roam the classroom to check for student understanding. On average, Page took two pages of longhand notes per observed lesson. The notes typically contained four to six comments about student and teacher actions. For other school staff (special education teachers or guidance counselors), Page observed similar amounts of time in the relevant contexts of practice and noted the flow of events.

We found that Page's observation notes focused mainly on the teacher's use of questioning, student behavior, and pacing and coherence of the lesson. Although Page noted that her observation method was rooted in her experience with Stillwater's previous teacher evaluation system, she also reported using the Instruction component of the Danielson framework and issues identified by the teacher in the TPGP preobservation discussion to guide her observation notes.

The choice of classroom lesson to be observed was also left unstructured by the TPGP. The tradition at Stillwater was to let teachers choose the lesson. From our interviews, we found that most teachers made their choices based on (a) their ability to exhibit proficiency in the SAR domains and (b) the potential of receiving relevant feedback from Page. One teacher noted how TPGP evaluation process afforded both tasks: "So I took a chance going in because how I look at the evaluations is it's the feedback. And if I use that lesson again how can I tweak it to make it better." Another teacher commented,

> There would have been days I would have not let her come in just because it would be boring. Like a test day for example, or even a day when they were presenting dialogue or something. It would be fun but you wouldn't see me teach. So there would be days I would say don't come in now.

One veteran teacher felt little risk in Page's visit and welcomed another perspective on his teaching, whereas another chose a routine Spanish recitation lesson because it fit conveniently into her schedule. Allowing teachers to select their lessons let teachers set the evaluation agenda. Both Page and the Baxter teachers opted out of evaluating either particularly difficult lessons or focusing the evaluation on lessons linked to key instructional initiatives that might provide feedback on Baxter's instructional priorities.

Micro Task 4: SAR Preparation—Calculating and Qualifying Teacher Quality

Within several days of the observation, Page began the 2-hour process of SAR write up for each individual teacher. The official record of teacher

performance at Stillwater, the SAR consists of two parts: a rating form according to the dimensions of the Danielson rubric and a narrative section for evaluators to provide a context and evidence for their rating. The SAR provides no supporting documents to guide evaluators on rating teachers or selecting appropriate evidence.

Again, Page relied on her reading of the SAR content and her experience as a teacher and an administrator to develop the following procedure: (a) read through the teachers' personnel file and observation notes, (b) compare her evidence (observation notes, personnel file, or memories of good/bad incidents) to the rubrics described in each SAR element, (c) begin with a "proficient" rating in each element and decide whether to move the teacher down to "basic" or up to "distinguished based on the evidence," and (d) write the SAR narrative by referencing her view of the teachers' role in school leadership, value to the school, growth/development during the school year, and SAR scores to explain ratings and summarize the teacher's contributions.

As the procedure shows, Page reached beyond the observation session to include past interactions with teachers and parent/student reports. This excerpt from a postobservation conference demonstrates the range of evidence Page called on:

> You have done that all along with kids . . . even when you get frustrated with some of those kids. *I know because I have sat in parent conferences with you.* I've always sensed that the student feels very much valued by you even when you are pointing out how they could improve. *And I have gotten feedback from parents, too*, that they are happy. . . . You set high expectations with kids. *I can tell in our conversation earlier, just now*, that you find certain things just unacceptable and you are going to maintain that high standard with kids because it affects the whole school. (emphasis added)

Teachers realized Page's data collection methods were limited and that ultimately her discretion drove evaluation scores. One teacher wondered about the adequacy the data Page uses for evaluation:

> I knew she [Page], she's always had this opinion of me that I can do no wrong. I don't know what that stems from. I think it's because we worked together for 20 years when she was teaching and in her time as assistant principal and now as principal she doesn't get any negative comments from students or parents. I'm not saying that's not good data. Its important data but it's not the whole picture. I think [Page] thinks I don't fail with kids but I do.

Page placed considerable emphasis on the SAR narrative as a way of making sense of her evaluation process. She spent 1 to 2 hours crafting each summary because she believed teachers view the narrative as the most important

aspect of the TPGP. The statements were typically 200 to 300 words and addressed four basic issues: (a) the preobservation conference purposes and the observation; (b) a general discussion of teachers' development, including areas of distinguished work; (c) suggestions for improvement; and (d) comments about the teacher's value to the Baxter community.

Page used the narratives to help connect the rating process with the teacher's goals and lesson design. Page felt that the SAR narrative allowed her to enter her comments about the teacher's quality into organizational memory.

In this example of a typical narrative, we can see how Page integrated the features of the TPGP artifact together with her perception of the teacher's role in the school. Page set the stage positively by linking the classroom observation to the Danielson framework:

> I had the pleasure to observe Ms. Reston in her third period, seventh-grade class. The students were involved in a number of engaging activities related to practicing the use of positive/negative and female/male adjectives and using the forms of the verb "to be" and subject pronouns correctly. Ms. Reston masterfully engaged students in learning. As Charlotte Danielson states in her book, "Engaging students in learning is the *raison d'etre* of education." Reston's practice in the classroom is a prime example.

Page shifted to a discussion of how Ms. Reston conducted her teaching by blending praise with specific details from the classroom observation.

> Ms. Reston captured the interest and attention of her students by guiding them through a variety of visual, verbal, and written exercises that were highly engaging. Her students understood exactly what skills they were reviewing or being introduced to before instruction. Ms. Reston's use of flash cards, oral drill, cooperative pair/share activity, the overhead, board work with laminated cards, and final application on a homework assignment all presented clear introduction and closure. Ms. Reston's use of materials, pacing, and lesson structure ensured a highly successful lesson.

Page linked the content of the preconference observation form to the narrative to show the connection between the stages of the evaluation process and to praise Ms. Reston's understanding of student needs:

During our preconference discussion, Ms. Reston asked that I watch for involvement and participation of particular students. Such concern and sensitivity to the individual needs of students is the mark of an outstanding teacher. Knowing these students, I was most interested to observe them, their responses, and Ms. Reston's awareness and interactions with them. I found both students attentive throughout the lesson, a credit to Ms. Reston's skill.

Page concluded the narrative by summarizing Ms. Reston's contributions to the school and to the district by noting Ms. Reston's status in the Stillwater community:

> Ms. Reston's contributions to the district as a whole are ongoing. Her hard work and dedication are noticed beyond Baxter Middle School. Recently [District Curriculum Director] Mr. Carlson sought me out to comment about his admiration for Ms. Reston's technological work on our district web site. Her commitment to the Stillwater District is evidenced not only by her excellent performance at Baxter but also by her professional activities that extend beyond the school day as evidenced by Mr. Carlson. Ms. Reston is an excellent teacher whose dedication and leadership contribute substantially to Baxter and to the district as a whole.

Page's narratives appeared designed to make sense of the evaluation process by situating her comments within the school and evaluation context.

Micro Task 5: Postobservation Conference

The postobservation conference provided an opportunity for Page to discuss the teacher's professional goals, self-assessment, and the SAR. The average postobservation conferences lasted an hour. Page's postobservation conference followed these steps: (a) check in on personal or school activities/happenings; (b) make general comments about her overall impressions of the teacher's work; (c) review teacher self-assessment (if available) and ask how the teacher thinks the lesson went; (d) reflect on the observed lesson by asking specific questions about the lesson; (e) report and explain aspects of SAR ratings, explained "basic" ratings, areas for improvement, or areas where SAR ratings differ from self-reflection ratings; (f) read, in full, the SAR narrative aloud to the teacher; (g) elicit teacher question/comments about the evaluation; (h) sign all appropriate forms. Should the teacher and principal agree on the substance of the SAR, the conference would conclude with each party signing a document to certify the agreement to be filed in the teacher's personnel folder.

Our analysis showed that Page varied the central message of the evaluation by sometimes emphasizing the teacher professional development plan, other times by using the observation ratings to suggest changes in teacher practice. With the probationary teachers, Page usually inserted additional goals into the conference to suggest further development.

Page read the SAR narrative verbatim during all postobservation conferences. Page explained that teachers then had the opportunity to raise questions and understand why ratings or statements were chosen. She paused to

relate how the critical suggestions for improvement should be understood in terms of the teachers' professional growth and goal statements. Page felt that reading the SAR narrative created a legitimate space for providing criticism while preserving her relationship with teachers. Following the narrative reading, Page reviewed the SAR rubric scores and commented on particularly high or low scores. With probationary teachers, Page pointed toward lower scores as areas of improvement and suggested how teachers could to tap into school/district resources for help.

With veteran teachers, Page tended to counterbalance the often-critical self-ratings provided by the teacher. One veteran teacher, for example, used the self-evaluation form to identify areas of her own practice she perceived as needing improvement. Instead of confirming these areas for improvement, Page instead suggested that the school, not the teacher, was responsible:

> Teacher: I had a couple basics [on my self-reflection form].
>
> Page: I thought you did too, and I was looking for them. I couldn't find them. [Page pulls the forms from the personnel file, turning them open. The teacher points to them, smiling.] There they are. That's right, under "resources for teaching" and "resources for students." And then how I'd expressed, you know, that I see you as just a very talented teacher reaching, reaching the kids on many many levels. And that I see you as being quite resourceful.
>
> Teacher: I'm not saying that I am not resourceful. I think the analogy for you there is that, for example, we have a school psychologist, a school social worker, and two counselors and invariably a kid comes to me with a problem and I don't know which person to send them to. And I should know that, after twenty-seven years in this field, I should know that.
>
> Page: And part of that problem will, hopefully, be solved next year because you are not unique to that. I think it is more of a school-based problem that we have certain people as resources, but they are part-time teachers.

Page seemed to deflect the teacher's low self-ratings with comments about schoolwide issues. Page recognized the teacher as a leader in the school and seemed reluctant to include negative ratings in her SAR.

In this case, the strategy to mediate the negative ratings with the teacher's role in the school may have backfired. In a later interview, the teacher stated that the postobservation conference comments made it look like she "walked on water." The teacher said she was somewhat disappointed that Page did not see her as she truly was, a teacher who could continue to improve. When asked about the SAR, the teacher said,

> It just focuses on things I hadn't thought of. . . . I am very in tune with where the kids are and how I'm connecting with the kids. But there were areas where like

> [Page] didn't like I scored myself low in some areas. But like resources in the school—I hadn't even thought of that. I'm so zeroed in on my classroom I don't tend to think beyond my classroom.

Page's rhetorical strategy to mute the teacher's self-criticism demonstrates the role of evaluator discretion. By shifting the evaluation process away from critique, Page emphasized her role as personnel manager in each instance. She judged that it was more important to emphasize the experienced teacher's value to the school community than to engage in substantive discussion about the teacher's critical self-ratings. Although this conversation may have been a missed opportunity for a substantive conversation about practice, Page used her discretion to steer the conversation away from the lower teacher self-scores, reasoning that lower scores at this stage of an accomplished teachers' career could alarm the central office or others external to the school community.

Page also used the postobservation conference as a structured occasion to interact with teachers about a wide range of topics not necessarily related to the classroom observation. In our content analysis of the postobservation discussions, we found that nearly half (48%) of the conference time was spent on checking in on personal or school activities/happenings. This example shows how Page shifts from the observation to discuss the issues involved with a particular student:

> Page: For showing professionalism, you have always stepped up whenever we have needed it for the school. And in your service to kids, well, you have had some kids with a lot of needs. And you have always been there to help them. And we should actually talk about one.
> Teacher: I've been hearing rumors.
> Page: I should fill you in. But he [the student] wants to stay at [the alternative program]. I should fill [the special education teacher] in too. We will need to have a team meeting. But that is going to be an important thing for us to do. But you have been doing a nice job though here at Baxter . . . .

The conversation continued for 7 more minutes about strategies for successfully helping this student. For Page, the postobservation conferences provide a structured opportunity to address emergent administrative and human resource issues to maintain the school's professional community.

The Cognitive System for Teacher Evaluation

The cognitive system at Baxter consisted of a complex network of artifacts and actors. Although our analysis focused mainly on the artifacts

directly addressed in the teacher evaluation process, we also found other artifacts that influenced evaluation in the broader cognitive system. One artifact stood out: the comprehensive school reform plan, experiential learning (EL). The EL reform model included a package of artifacts to guide instruction, assessment, professional learning, and school governance. EL highlighted the importance of learning expeditions, collaborative, interdisciplinary projects that resulted in authentic products for real audiences. The school had invested considerable time and resources over the previous 4 years in EL. Grade-level teacher teams met regularly with EL consultants to designed expeditions that integrated learning across subject areas.

The TPGP artifact was adopted at a critical time for EL at Baxter. Page felt that promoting EL at Baxter required developing a strong professional learning community. The initial EL reform design required professional development for the whole school but made the development of EL curriculum development projects voluntary for teachers. Page believed that the trusting, collaborative professional communities of teachers and administrators were needed for building support for EL. Page felt that there were still a number of teachers who did not yet buy into the EL design. A staff survey revealed that although a majority of the Baxter staff initially approved the EL initiative, only 25% of teachers had actually aligned their teaching practices with EL after 6 years of reform. She felt that if she could get a few more teachers on board with EL, the school might reach a tipping point at which widespread use would become inevitable. Page thus minimized the time and resource costs of TPGP due to the work yet to be done for EL, reasoning the demands of TPGP might threaten a staff already burdened with EL.

The influence of EL was also felt within the evaluation process itself. EL already included a formative evaluation emphasis for teacher work. Page decided to align early TPGP implementation with EL by playing up the similarities between formative aspects of EL and TPGP evaluation practices and playing down the summative evaluation requirements of TPGP that appeared contrary to the formative spirit of EL evaluation. Although Page felt that, in time, teachers would come to see how the summative features of the TPGP were consistent with EL, she decided to emphasize the goal-setting features consistent with EL in the pilot TPGP implementation. In the interest of promoting a coherent approach to instruction, Page often used EL examples to illustrate teacher's goals and classroom practice.

EL may have been the most prominent program at Baxter but was not the only artifact in the broader cognitive system that influenced TPGP implementation. We observed how the daily schedule, the prior evaluation system, the student support system, and the existing curriculum also shaped evaluation. These aggregated artifacts comprised the structural components of the

cognitive system evaluation at Baxter. Page used her discretion to mediate between features of this cognitive system, features of the TPGP, and her perception of teacher needs to shape her use of the Stillwater evaluation system.

DISCUSSION

The Baxter case includes many familiar features already documented in both practitioner experience and research. The case relates the familiar story that the context influences implementation and can distort policy signals (e.g., Talbert & McLaughlin, 1993). Our distributed cognition analysis, however, helps us move beyond the general observation that the context matters to show which aspects of context matter when and how practitioner discretion distorts signals. The case displays the range of artifacts that actually guided evaluation practice and demonstrated how the school's existing cognitive system influenced evaluation. Karen Page needed to compare the benefits and costs of implementing artifact features to her existing commitments to people and programs within the school, and she had to find balance in her role as evaluator (or referee) and human resource developer (or coach). In the following sections, we expand on how our distributed cognition analysis shows how the context influenced artifact design and use in three key areas of the cognitive system. First, we use the idea of the cognitive system to explain how the policy design relied on underlying artifacts. We then focus on the intended and actual role of discretion in evaluation, then consider the relation of routines to evaluation practices.

Layers of the Cognitive System

We opened our argument with the suggestion that policy development and implementation studies need a better way to understand how practitioners use policies and how new policies enter and enliven existing systems of practice. Our analysis uncovered several aspects of the cognitive system for teacher evaluation at Baxter.

The TPGP design initially provided a series of district-designed artifacts to guide the evaluation process. The district design rationale suggested that the artifacts supplied by the TPGP would structure the evaluation process for both teachers and evaluators. The vision seemed to assume that the evaluation process for teachers could be structured with the same artifacts used by leaders and that the new evaluation system could be implemented in isolation from the rest of school context.

Our analysis showed that the teacher evaluation task revealed aspects of the school's working cognitive system that supplemented the constraints provided by the district tools. Although TPGP artifacts structured the evaluation process for teachers, the artifacts proved inadequate for evaluator efforts to collect evidence or to mediate tensions between teacher-selected goals and the TPGP framework. Page introduced other artifacts to guide her classroom observation practice and her postobservation conferences that enabled her to structure processes left unstructured. The absence of TPGP artifacts to guide practice resulted in an evaluation process as attentive to other goals, such as organizational maintenance and ongoing communication with teachers, as it was focused on substantive issues of improving teaching and learning. Page attempted to integrate the instrument onto her school's existing cognitive system. She saw the TPGP artifact in relation to her previous evaluation efforts, her human resource development efforts, and school reform. The wider cognitive system at Baxter also included artifacts that influenced Page's evaluation practice. We saw how the EL program, for example, shaped how Page presented the TPGP program to her teachers and influenced her emphasis of certain TPGP features over others so that she could gain support for TPGP without sacrificing the ongoing instructional reform agenda. To reduce the cognitive load of implementing a new program, and the load on her staff, she attempted to offload some of process onto already robust features of the existing cognitive system.

The district policy development process could not possibly be expected to address the range of potentially relevant local artifacts. Cataloguing the artifacts that could shape implementation would be a daunting task of questionable value, in part because of the sheer number of artifacts that could influence practice and in part because artifacts would have different influences in different contexts. The task of navigating the local context is rightfully given to local practitioners. However, policy makers and practitioners could use distributed cognition analyses to understand which artifacts influence the implementation of new programs to shape artifact and systems design.

Our distributed cognition analyses point toward the junctions in policy use in which practitioners need better tools to guide practice. At Baxter, Karen Page was left to her own devices to negotiate the trade-offs between old and new practices. Closer attention to how contexts influence practice could have policy designers design evaluation tools to effect intended changes in the context of existing initiatives. The distributed cognition analysis provides a lens for the adaptive processes associated with policy implementation. This new lens enables policy development and implementation research to move beyond discussion of policy fidelity to design artifacts for

use in local cognitive systems that allow practitioners to rely on familiar contexts to improve teaching and learning.

Evaluator Discretion

Focusing on the cognitive system as the unit of analysis reveals the importance of discretion as the key cognitive activity of teacher evaluators. Discretion here refers to the actor's power to use judgment to determine a course of action within the perceived constraints of a situation. Hambrick and Finkelstein (1987) used the concept of managerial discretion to describe the latitude managers have to influence organizational outcomes. Discretion is a more nuanced version of volition or intention that refers to how choices and decisions are bounded by the perceived constraints of a given situation. Policy makers both constrain and depend on managerial discretion through the design of artifact features. Policies are designed to constrain practitioners' behavior to produce intended practices and outcomes. In this sense, policies aim to inhibit local discretion through artifact features designed to guide appropriate action. However, policies also rely on practitioner discretion to adjust policy demands to local circumstances or to fill in gaps left unspecified by policy design.

We might conclude at this point with the simple and long-standing observation that the quality of evaluation depends on the discretion of the evaluator. A distinguishing feature of a distributed cognition analysis, however, is the ability to make cognitive activity, such as managerial discretion, accessible through the operation of a cognitive system. In ordinary cognitive analysis, researchers must infer the operation of cognitive processes in individuals, such as intentionality or problem setting, from observed behaviors. In a distributed cognition analysis, artifact design and use externalizes cognitive activity, and tasks demonstrate how actors access the memory and resources of the cognitive system through the selection (or suppression) of artifact features. Discretion is thus made visible (and accessible) in a cognitive system through how evaluators select from artifact features in tasks. Here we address two aspects of discretion in the TPGP evaluation practices: how discretion serves as a core cognitive component of leadership expertise and how discretion can be taught in a cognitive system.

Discretion and Leadership Expertise

Managerial discretion explains how managers can have different records of success in organizations with comparable capacity (Hambrick & Finkelstein 1987). Discretion helps explain why, in the same systems, some

leaders see opportunities for change when others see ironclad constraints on action. We argue that discretion plays a dual role of guiding both problem setting and problem solving: the actor's read of the situation determines the kind of problem to be solved, and the actor's understanding of the system capacity shapes the proposed solution. Discretion here refers to the iterative ability to adapt new artifact features to existing commitments while at the same time reevaluating existing commitments in light of the capacities of new artifact features. In other words, the exercise of discretion provides a key component for understanding leadership expertise. Bereiter and Scaradamalia's (1993) expertise research uses the concept of progressive problem solving to explain how experts continually challenge and stretch their existing cognitive frameworks by seeking out new problems. In the field of school leadership research, Leithwood, Begley, and Cousins (1992) explained progressive problem solving in terms of the leader's cognitive flexibility to respond to the possibilities of emergent circumstances while at the same time balancing a commitment to core principals and avoiding the blind alleys that come from inflexible adherence to prior understanding. In a distributed cognition analysis, cognitive flexibility is displayed through the actor's discretion about which aspects of the cognitive system to emphasize and which to play down in a given situation.

Page's expertise was displayed in her ability to adjust the features of the artifact to her perception of the needs of her teachers and the school community. Page's evaluation practice reflected her commitment to preserve the relational trust necessary for maintaining existing initiates in the school. She balanced issues of tenure, expectations, and the position of the teacher in the school to select which artifact features to emphasize. For a veteran teacher, Page allowed the goal statement to dictate the evaluation report. The teacher was encouraged to talk about the goals she set for herself in the evaluation session. For a probationary teacher, Page reviewed the areas for improvement from the last evaluation and used the SAR to suggest new areas to work on. On the few occasions in which she criticized her teachers, she used the narrative aspect of the SAR and the postobservation conference to carefully situate her critical comments in a supportive, formative context. Our analysis showed how, even when teachers tried to push the more critical aspects of the evaluation framework by providing harsh ratings for their own practice, Page sought to restore exemplary ratings by assuring teachers that the fault was with the school program, not with the teacher's practice. This suggests Page used her discretion to mediate an apparent trade-off between professional support and development and summative evaluation.

The importance of teacher evaluator discretion changes based on the intentions built into of the policy artifacts. Policies that aim for standardized

practice seek to restrict the range of discretion, whereas policies designed to enhance the range of tools available to local practitioners depend on discretion. From the perspective of TPGP developers, Page's discretion to downplay the summative and critical aspects of the district evaluation artifact may be read as an implementation failure. Particularly from the perspective of the district members of the design team, her decisions about when to press for the priority of local artifacts conflicted with implementing the more demanding standards for teaching built into the TPGP framework. However, from the perspective of practitioner expertise, Page's ability to selectively implement artifact features on the fly signaled neither a lack of courage nor a lack of ability to enforce the harsher standards of the framework. She used the tools provided by the TPGP to work toward several organizational goals. Page recognized the fragility of the faculty consensus required to maintain existing organizational initiatives. Principals interested in reform-based practices work need to maintain the relational trust necessary for teachers to abandon their autonomy and try something new (Bryk & Schneider, 2002). Our analysis revealed when Page used her discretion to emphasize organizational goals over new artifact features to help her teachers accept and participate in the TPGP. The distributed cognition analysis established a framework to access discretion as a form of leadership expertise in action.

Learning Discretion

The kinds of discretion guided by policy and used in practice need not be opposed to one another. Cohen and Hill (2001) suggested that successful policy implementation is marked by opportunities for implementers to learn. In the TPGP example, both designers and practitioners need opportunities to learn from each other about (a) how policies are intended to change practices and (b) how practices need to inform policy development. A distributed cognition analysis points to a how an iterative design process might help policy designers build tools to guide local discretion in intended directions. An iterative design approach sees artifact development as a work in process and uses a cyclic metaphor to continually revisit initial design assumptions. A distributed cognition analysis provides the foundation for establishing an iterative design process that could take note of the how (and where) evaluators introduced artifacts into the TPGP process to anticipate where new support tools should be built. An iterative approach to artifact design would also help to address the tendency of evaluation practices to drift, as it were, to organizational maintenance issues. Seifert and Hutchins (1992) suggested that "it is much more difficult to design for learning than for system performance" (p. 97). The artifacts and practices that Page introduced to supplement the

evaluation process pointed toward opportunities for the district team to refine the TPGP artifact. Unfortunately, Page's reliance on familiar artifacts served to undercut some of the more challenging features of the TPGP framework. To direct wholesale changes in Stillwater teacher evaluation practices, the designers needed to understand how to reshape teacher evaluator practices as they evolved. An iterative design process might help designers how to help evaluators learn to use discretion consistent with the central TPGP goals.

Our analysis showed, for example, that the TPGP artifacts were silent about where the summative aspects of the evaluation process, such as the selection of classroom evidence for justifying ratings or structures to guide narrative writing that would connect feedback to ongoing school instructional goals, could best be brought into play. Learning to select appropriate evidence to justify observation ratings is a critical aspect of any evaluation process. Nelson and Sassi (2000) argued that the critical aspect for successful evaluation practice is how evaluators develop an "eye" for the relevant aspects of teaching. Once the district team realized that training to develop an "eye" for teaching was an important design outcome, they could begin to consider a range of artifacts to help evaluators to recognize good questioning and to notice who is shouldering the burden of thinking in classrooms to help for evaluators to "see" teaching practice in new ways. In the postobservation conferences, artifacts could be designed to help evaluators organize evidence and structure conversations geared toward improvement. Acquiring an eye for supporting good teaching also involves building a repertoire of skills and techniques to provide the appropriate support. When the development of these capacities is not addressed at the artifact design level, it is little wonder that many evaluators would emphasize familiar goals over the undeveloped capacities necessary to successfully implement challenging artifact features. Designers interested in making policies that influence practice in intended ways need to attend to the tasks and artifacts necessary to help practitioners use their discretion and act on the right things.

Evaluation and Routines

In this final section, we consider how the concept of routine relates to teacher evaluation. Routines play a central role in Hutchins's distributed cognition analyses. The macro tasks Hutchins considers, such as ship navigation and piloting, are composed of a number of micro tasks that together compose routines. The concept of routine has also received considerable attention in organizational theory and sociology. Routines have been described as building blocks of organizations (Cyert & March, 1963) that explain how practices persist over time. Routines establish patterns of interaction between

cognitive schema (Ashforth & Fried, 1988; Schank & Abelson, 1977) and the expected "performance program" of the organization (March & Simon, 1958, p. 142). Routines are paths established by trial and error through complex situations that anticipate the regular obstacles and provide standardized access to useful artifacts. Giddens (1984) argued that routines "represent the institutionalized features of social systems" (p. 86 as quoted in Pentland & Reuter, 1994). However, Giddens (1984) suggested that a routine is more than a static construct. Rather, "the routinized character of most social activity is something that has to be 'worked at continually'" (p. 86). In other words, a routine represents a developmental achievement. The aim of a navigational routine, for example, is to develop a standard operating procedure (SOP) to reduce the need for actors to improvise their way through complex tasks. SOPs constrain the variability of a complex environment by directing simple tasks so that actors can use their discretion to focus on the unpredictable.

What role do routines play in the analysis of teacher evaluation practices? Here, our interpretation of the distributed cognition framework must also be seen in a developmental context. In mature systems with well-established SOPs, new artifacts are used to predictably reduce the range of practitioner discretion. However, in emergent systems without SOPs or in transition from one SOP to another, the need for practitioner discretion is magnified rather than reduced. In our case, the TPGP artifact was newly implemented in the context of tangled prior evaluation practices. These prior practices were often accidental (unannounced visits to classrooms for nonevaluation purposes), ineffective (prior district evaluation practices), or incidental (evaluation through participation in instructional initiatives such as EL). District designers sought to establish a new evaluation routine with the TPGP artifact sequence. Introducing a new artifact increased the discretionary burden by forcing evaluators to fit the new artifacts in the context of their prior evaluation experience and in the existing cognitive system. Introducing the new TPGP artifacts into Baxter's rich situation of practice meant that new routines had to be established to counter the organizational inertia of the existing routines. The implementation of the new artifacts thus made the situation less predictable, widening the scope of evaluator discretion to help make sense of the new practices in terms of the old. In the absence of a yet-to-be-established routine, Principal Page relied on her discretion to make sense of the TPGP for her school.

Does this analysis suggest the development of a teacher evaluation SOP is either a desirable or an appropriate outcome? Policy makers pushing for a standardized, predictable measure of teacher performance through a knowledge- and skill-based evaluation artifact might hope that evaluation routines

will be developed to reduce local variation. Reducing the variation in implementation would mean less reliance on evaluator discretion and more reliance on predictable and shared procedures. The emergence of SOP would lead to more consistent measures of teaching, which in turn would provide a more stable measure of teaching performance and reliable data for system leaders to better allocate resources. From this perspective, a reliable SOP is a desirable outcome for teacher evaluation artifacts.

Practitioners, on the other hand, might prefer to retain considerable discretion in evaluation practice because they do not see the artifact set in isolation but as a part of a broader system. Hutchins's broader system was the airplane, and Page's system was her school. Practitioners seldom have the luxury of implementing artifacts on a clean slate. There are always prior and competing artifacts already in place, and traditional routines shaped by inherited artifacts provide powerful constraints on new practices. Following Perrow's (1984) insight that intelligent operators are required to intervene even in well-designed systems with redundant error-checking mechanisms, implementing new evaluation artifacts would require practitioners to make continuous corrections between the features of the new and existing artifacts. A new evaluation program might establish a macro-level SOP, but negotiating the details of teacher needs, traditions of practice, and institutional requirements will always rely on evaluator discretion. As Dornbusch and Scott (1975) explained,

> Appraisal is seldom a mechanical procedure . . . appraising a task requires knowledge of extenuating circumstances. Such information is of critical importance in determining what, if any, message is to be communicated to the performer concerning the quality of his or her task performance. (p. 143)

Reducing these discretionary aspects of evaluation to routines could result in systems very much like the practices a program such as TPGP was designed to replace: empty, formalized practices that provide little helpful feedback for teachers or leaders. In the TPGP analysis, the trade-off between predictable outcomes and adaptability rests on the evaluator's ability to use discretion consistently with the central artifact features. Although Hutchins's analyses of distributed cognition do not eliminate actor discretion from practice, the development of reliable SOPs on the ship reduced the need for discretion to seek out unanticipated inputs. Our analysis shows that the design of TPGP made each teacher a source of unanticipated input and that the implementation of the Baxter evaluation system required the evaluator to continually use discretion to fit policy goals with organizational needs.

CONCLUSION

Although it has long been recognized that the local context influenced policy implementation, our distributed cognition analysis of teacher evaluation practice demonstrates how the local context shapes practice. A traditional distributed cognition analysis, as described by Hutchins, demonstrates how tasks flow through existing systems of tools and actors. Our distributed cognition analysis examined how the task of teacher evaluation in schools was mediated by a district-deigned teacher evaluation system. This case provides a rich illustration of how artifact design contributes to (and relies on) organizational practice and local discretion. Our analysis showed that the actual cognitive system for evaluation was far more complicated than envisioned by the designers and how evaluators need to rely on their discretion to resolve the conflicting artifact features in practice. The artifact design team engaged in a collaborative design process that brought stakeholders together to reshape the evaluation artifact. But the collaborative design of TPGP also had limitations. The Stillwater design process failed to take a clear stand on the balance between summative or formative features. Teachers and principals involved in the design process recognized the clear differences between summative and formative evaluation and disagreed how they could be incorporated into the same process. Incorporating both functions into the Stillwater TPGP pushed evaluators to use their discretion to negotiate the tension between the summative and formative policy features.

Identifying the relevant aspects of the cognitive system points to the key challenge practitioners need to address in implementing new artifacts in schools. A distributed cognition analysis helps to make sense of the evocative but generic concept of "context" to pinpoint just where evaluators feel the need to make trade-offs between new artifact features and existing commitments. We hope that this article showed the promise of a distributed cognition analysis for opening the black box of practice and for highlighting how policy makers can better design artifacts to support practitioner learning.

The study also points to limitations of distributed cognition analyses. Focusing on tasks and artifacts can sidestep the social dimensions of change, overlook the motivations of people to engage in or resist change, and ignore the micro-political structures intended to protect teaching practice from external inspection and intervention. These issues remain as key sociopolitical aspects of any effort to change teaching and learning practices. Still, no analytic framework can address all aspects of a complex problem. Our sample application of the distributed cognition framework revealed the interplay of tasks, artifacts, and actors in teacher evaluation. Future applications of the distributed cognition framework to chronic school problems,

including but not limited to teacher evaluation, could incorporate more perspectives (e.g., teachers and district leaders) or more cases of evaluation practice to provide a richer, more nuanced view of practice. This case represented our preliminary effort to use a distributed cognition perspective to reveal how practitioners make sense of new practices in terms of the existing system. The tacit connections in the system, once made explicit, reveal the bottlenecks in implementation that both policy designers and policy users can use to have a better chance to improve teaching and learning in our schools.

APPENDIX
Danielson (1996) Framework Outline

COMPONENTS OF PROFESSIONAL PRACTICE

Domain 1: Planning and Preparation

1a: Demonstrating Knowledge of Content and Pedagogy
- Knowledge of content
- Knowledge of prerequisite relationships
- Knowledge of content-related pedagogy

1b: Demonstrating Knowledge of Students
- Knowledge of characteristics of age group
- Knowledge of students' varied approaches to learning
- Knowledge of students' skills and knowledge
- Knowledge of students' interests and cultural heritage

1c: Selecting Instructional Goals
- Value
- Clarity
- Suitability for diverse students
- Balance

1d: Demonstrating Knowledge of Resources
- Resources for teaching
- Resources for students

1e: Designing Coherent Instruction
- Learning activities
- Instructional materials and resources
- Instructional groups
- Lesson and unit structure

1f: Assessing Student Learning
  Congruence with instructional goals
  Criteria and standards
  Use for planning

Domain 2: The Classroom Environment

2a: Creating an Environment of Respect and Rapport
  Teacher interaction with students
  Student interaction
2b: Establishing a Culture for Learning
  Importance of content
  Student pride in work
  Expectations for learning and achievement
2c: Managing Classroom Procedures
  Management of instructional groups
  Management of transitions
  Management of materials and supplies
  Performance of noninstructional duties
  Supervision of volunteers and paraprofessionals
2d: Managing Student Behavior
  Expectations
  Monitoring student behavior
  Response to student misbehavior
2e: Organizing Physical Space
  Safety and arrangement of furniture
  Accessibility to learning and use of physical resources

Domain 3: Instruction

3a: Communicating Clearly and Accurately
  Directions and procedures
  Oral and written language
3b: Using Questioning and Discussion Techniques
  Quality of questions
  Discussion techniques
  Student participation
3c: Engaging Students in Learning
  Representation of content
  Activities and assignments
  Grouping of students
  Instructional materials and resources
  Structure and pacing

3d: Providing Feedback to Students
  Quality: accurate, substantive, constructive, and specific
  Timeliness
3e: Demonstrating Flexibility and Responsiveness
  Lesson adjustment
  Response to students
  Persistence

Domain 4: Professional Responsibilities

4a: Reflecting on Teaching
  Accuracy
  Use in future teaching
4b: Maintaining Accurate Records
  Student completion of assignments
  Student progress in learning
  Noninstructional records
4c: Communicating With Families
  Information about the instructional program
  Information about individual students
  Engagement of families in the instructional program
4d: Contributing to the School and District
  Relationships with colleagues
  Service to the school
  Participation in school and district projects
4e: Growing and Developing Professionally
  Enhancement of content knowledge and pedagogical skill
  Service to the profession
4f: Showing Professionalism
  Service to students
  Advocacy
  Decision making

REFERENCES

Aristotle. (1941). On the soul. In R. McKeon (Ed.), *The basic works of Aristotle* (pp. 535-606). New York: Random House.

Ashforth, B. E., & Fried, Y. (1988). The mindlessness of organizational behaviors. *Human Relations*, *41*, 305-329.

Bereiter, C., & M. Scardamalia. (1993). *Surpassing ourselves: An inquiry into the nature and implications of expertise*. Chicago: Open Court.

Brown, A. L., & Campione, J. C. (1996). Psychological theory and the design of innovative learning environments: On procedures, principles and systems. In L. Schauble & R. Glaser (Eds.),

Innovations in learning: New environments for education (pp. 289-325). Mahwah, NJ: Lawrence Erlbaum.

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

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

Cole, M., Engeström, Y., & Vasquez, O. (1997). *Mind, culture, and activity.* Cambridge, UK: Cambridge University Press.

Cyert, R. M., & March, J. G. (1963). *A behavioral theory of the firm*. New York: Prentice Hall.

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

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

Dornbusch, S. M., & Scott. W. R. (1975). *Evaluation and the exercise of authority.* San Francisco: Jossey-Bass.

Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real-world laboratories. In R. J. Sternberg & J. E. Davidson (Eds.), *The nature of insight* (pp. 365-395). Cambridge, MA: MIT Press.

Goodwin, C. (1995). Seeing in depth. *Social Studies of Science*, *25*, 237-274.

Giddens, A. (1984). *The constitution of society.* Berkeley: University of California Press.

Halverson, C. A. (1995). *Inside the cognitive workplace: New technology and air traffic control*. Unpublished doctoral dissertation, University of California, San Diego.

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

Halverson, R. (2003). Systems of practice: How leaders use artifacts to create professional community in schools. *Educational Policy and Analysis Archives*, *11*, 37. Retrieved March 23, 2005, from http://epaa.asu.edu/epaa/v11n37/

Halverson, R. (2004). Accessing, documenting and communicating practical wisdom: The *phronesis* of school leadership practice. *American Journal of Education 111*, 90-121.

Halverson, R., & Clifford, M. (2004). *How the situation of practice shapes the implementation of new policies in schools*. Paper presented at the American Educational Research Association annual meeting, San Diego, CA.

Halverson, R., Kelley, C., & Kimball, S. (2004). Implementing teacher evaluation systems: How principals make sense of complex artifacts to shape local instructional practice. In C. Miskel and W. Hoy (Eds.), *Theory and research in educational administration* (Vol. 3, pp. 66-90). Greenwich, CT: Information Age Press.

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

Hambrick, D. C., & Finkelstein, S. (1987). Managerial discretion: A bridge between polar views of organizational outcomes. In L. L. Cummings & B. M. Staw (Eds.), *Research in organizational behavior 9* (pp. 369-406). Greenwich, CT: JAI.

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 and Supervision*, *9*, 195-216.

Hollan, J., Hutchins, E., & Kirsch, D. (2000). Distributed cognition: Toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction*,

7, 174-196. Retrieved August 26, 2004, from http://citeseer.nj.nec.com/hollan00distributed.html

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

Hutchins, E. (1995b). How a cockpit remembers its speed. *Cognitive Science*, *19*, 265-288.

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*, 241-269.

Lave, J. (1988). *Cognition in practice: Mind, mathematics, and culture in everyday life*. New York: Cambridge University Press.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation.* New York: Cambridge University Press.

Leithwood, K., Begley, P. T., & Cousins, J. B. (1994). *Developing expert leadership for future schools*. London: Falmer Press.

Leont'ev, A. N. (1975). Activity, consciousness and personality (M. J. Hall, Trans.). Englewood Cliffs, NJ: Prentice Hall.

Leont'ev, A. N. (1981). *Problems of the development of the mind.* Moscow: Progress.

Lindblom, C. E. (1995). The science of muddling through. In S. Theodoulou & M. Cahn (Eds.), *Public policy: The essential readings* (p.113-127). Englewood Cliffs, NJ: Prentice Hall.

Lipsky, M. (1980). *Street-level bureaucracy: Dilemmas of the individual in public services*. New York: Russell Sage Foundation

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.

Mangin, M. (2004). *Distributed leadership and the culture of schools: Teacher leaders' strategies for gaining access to classrooms*. Paper presented at the 2004 annual meeting of the American Educational Research Association, San Diego, CA.

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

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

Merriam, S. B. (1998). *Qualitative research and case study applications in education*. San Francisco: Jossey-Bass.

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*, 193-212.

Natriello, G., Pallas, A., & McDill, E. L. (1990). *Schooling disadvantaged children: Racing against catastrophe*. New York: Teachers College Press.

Nelson, B. S., & Sassi, A. (2000). Shifting approaches to supervision: The case of mathematics supervision. *Educational Administration Quarterly*, *36*, 553-584

Neressian, N. J., Kurz-Milcke, E., Newstetter, W. C., & Davies, J. (2003). Research laboratories as evolving distributed cognitive systems. *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, 857-862.

Norman, D. A. (1991). Cognitive artifacts. In J. M. Carroll (Ed.), *Designing interaction: Psychology at the human-computer interface* (pp. 17-38). New York: Cambridge University Press.

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

Pentland, B. T., & Reuter, H. H. (1994). Organizational routines as grammars of action. *Administrative Science Quarterly*, *39*, 484-510.

Perkins, D. N. (1993). Person-plus: A distributed view of thinking and learning. In G. Salomon (Ed.), *Distributed cognitions: Psychological and educational considerations* (pp. 88-110). New York: Cambridge University Press.

Perrow, C. (1984). *Normal accidents: Living with high-risk technologies*. New York: Basic Books.

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

Rogoff, B. (1990). *Apprenticeship in thinking: Cognitive development in social context*. New York: Oxford University Press.

Salomon, G., & Perkins, D. (1993). *Distributed cognitions*. New York: Cambridge University Press.

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

Seifert, C. M., & Hutchins, E. L. (1992). Error as opportunity: Learning in a cooperative task. *Human Computer Interaction*, *7*, 409-435.

Sergiovanni, T., & Starratt, R. (1993). *Supervision: A redefinition*. New York: McGraw-Hill. **NOT CITED, DELETE?**

Simon, H. A. (1997). *Administrative behavior* (4th ed.). New York: Free Press.

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

Spillane, J. P., Halverson, R., & Diamond, J. B. (2004). Towards a theory of leadership practice: A distributed perspective. *Journal of Curriculum Studies*, *36*(1), 3-34.

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

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

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

Talbert, J., & McLaughlin, M. (1993). Understanding teaching in context. In D. Cohen, M. McLaughlin, & J. Talbert (Eds.), *Teaching for understanding: Challenges for policy and practice* (pp. 167-206). San Francisco: Jossey-Bass.

Vygotsky, L. S. (1978). *Mind in society: The development of the higher psychological processes* (A. Kozulin, Trans.). Cambridge, MA: Harvard University Press.

Wertsch, J. V. (1998). *Mind as action.* New York: Oxford University Press.

Zhang, J., & Norman, D. A. (1994). Representations in distributed cognitive tasks. *Cognitive Science*, *18*, 87-122.

Richard Halverson *is an assistant professor in the Educational Leadership and Policy Analysis Department at the University of Wisconsin–Madison. His work aims to bring the research methods and practices of the learning sciences to the world of educational leadership by exploring how to access, document, and communicate the practical wisdom of school leaders.*

Matthew Clifford *is an associate researcher at the Wisconsin Center for Educational Research. His research considers how and why school leaders—administrators, teachers, and others—create networks of support to sustain reform-oriented instructional practices, particularly in science and mathematics.*