Article

UC EA University Council for Educational Administration

Educational Administration Quarterly
2016, Vol. 52(2) 187–220

DOI: 10.1177/0013161X15617808
eaq.sagepub.com
SAGE

An Examination of the Benefits, Limitations, and Challenges of Conducting Randomized Experiments With Principals

Eric M. Camburn[1], Ellen Goldring[2], James Sebastian[3], Henry May[4], and Jason Huff[5]

Abstract

Purpose: The past decade has seen considerable debate about how to best evaluate the efficacy of educational improvement initiatives, and members of the educational leadership research community have entered the debate with great energy. Throughout this debate, the use of randomized experiments has been a particularly contentious subject. This study examines the potential benefits, limitations, and challenges involved in using experiments to evaluate professional development for principals. **Approach:** We present a case study of an experimental evaluation of a professional development program for principals. The case study is grounded in key themes in recent debates about the use of experiments in educational research, scholarship on challenges in conducting experiments, and experimental studies involving principals. **Setting and Sample:** The case study was conducted in an

[1]University of Wisconsin–Madison, WI, USA
[2]Vanderbilt University, Nashville, TN, USA
[3]University of Missouri–Columbia, MO, USA
[4]University of Delaware, Newark, DE, USA
[5]New Leaders for New Schools, New York, NY, USA

Corresponding Author:
Eric M. Camburn, University of Wisconsin–Madison, 1025 West Johnson St., 1186C, WI 53716, USA.
Email: camburn@wisc.edu

urban school district with 48 principals. **Implications for Research:** The experimental component of the study allowed us to form a trustworthy summary inference about whether or not a professional development program had an overall effect on principals. However, the experiment did not illuminate why or how the program failed to influence principal practice. Using descriptions of the intended curriculum for principals, professional development attendance records, and interview data, we developed an understanding of why the program failed to achieve its intended goals. Based on our experiences, we support continued advocacy of research designs that bring rich evidence to bear about causal mechanisms, implementation conditions, potential measures of delivery of and adherence to treatment protocols, and measures of participants' exposure to treatment.

Keywords
leadership, principals, professional development, experiment, leadership practice

Introduction

Principal preparation programs have come under intense scrutiny in recent years with critics charging that they often fail to prepare leaders for the realities of running modern schools (The American Association of Colleges for Teacher Education, 2001; Elmore, 2000; Levine, 2005). Professional development for practicing principals is viewed by some as a potential antidote to this problem (Peterson, 2002). Many states and districts appear to concur, devoting substantial resources to developing principals' leadership capacity (Augustine et al., 2009). There is growing belief that professional development for principals that has coherent, research-based content and that provides principals with authentic, problem-based, collaborative learning experiences can be effective in improving principal practice (Davis, Darling-Hammond, LaPointe, & Meyerson, 2005). Despite the optimism that some researchers, policymakers, and practitioners have about the potential benefits of principal development programs, evidence of the efficacy of principal professional development is scarce, and virtually none of it is from studies that permit strong causal inferences about program effects (LaPointe, Meyerson, & Darling-Hammond, 2006).

The past decade has seen considerable debate about how to best evaluate the efficacy of education improvement initiatives like professional

development for principals. The publication of the report *Scientific Research in Education* by the National Research Council has proven to be a significant milestone spurring a spirited conversation about the nature of inquiry and subsequent knowledge claims in education research. This debate also received substantial attention from the education leadership research community, including an edited volume requisitioned by the executive committee of the University Council for Educational Administration that was "centered on some of the premises and postulates of Scientific Research in Education" (English, 2007, p. xi–xii).

Throughout this debate, randomized experiments, the research strategy that is the focus of this study, has been a particularly contentious subject that has had both strong advocates (see, e.g., Boruch, 2002; Cook, 2002; Slavin, 2002) and critics (see, e.g., Howe, 2005). Advocates of so-called scientifically based research have called for greater use of randomized experiments to determine the efficacy of educational programs, and the federal government has promoted the use of experiments through the No Child Left Behind legislation and legislation authorizing the Institute of Education Sciences (Eisenhart & Towne, 2003; National Research Council, 2002).

This article examines the potential benefits, limitations, and challenges involved in using experiments to evaluate professional development for principals. Our examination is grounded in recent debates about the use of experiments in education research, scholarship on common challenges faced in implementing experiments in education, and a review of the state of experimental evidence on principals and principal development initiatives. We begin with a discussion of these three areas of research. We next present a case study of an experimental evaluation of a professional development program for principals to illustrate and examine key themes in the literatures. Specifically, the case study illustrates our mixed-methods approach to testing the effect of the professional development program on participants' leadership practice and presents the results of the experiment. The case study then illustrates how a range of factors affected program implementation and, subsequently, our ability to observe program effects. Finally, we reflect on the implications of the use of experiments for evaluating principal professional development programs and for future research by relating our experiences and results to ongoing debates on the nature of inquiry and knowledge claims in research on education leadership and the broader field of education research. We believe that this study is timely given that these debates remain unresolved and given the rarity with which this method has been used to study principals.

Background

Our study is framed by three lines of research. First, we situate our examination of the use of experiments for studying principal development programs within recent debates about the use of this method in the broad field of education research and in the subfield of research on education leadership. Second, we provide further context for the study by examining the state of experimental evidence on principals and principal development initiatives. Third, we discuss common challenges faced in implementing experiments in education, which gives us a basis for making sense of challenges we experienced and for reflecting on the potential benefits of experiments involving principals.

Debate About the Use of Randomized Experiments in Education

Since its inception, the field of education research has seen active debates about what constitutes valid evidence about education and its effects and about appropriate methods for producing such evidence. A recurrent theme has been whether and how education research might incorporate the methods of science (Condliffe Langeman, 2000). To varying degrees, these debates have made their way into the subfield of research on education leadership. As Crow (2007) points out, one of the original purposes of University Council for Educational Administration was to shift "educational administration from an anecdotal orientation to a more scientific one, leading to generalizations about organizations and leadership" (p. vii).

The 2002 publication of the report *Scientific Research in Education* by the National Research Council has proven to be a significant milestone in these debates. The stated approach of the report was to

> review and synthesize recent literature on the science and practice of scientific education research and consider how to support high quality science in a federal education research agency. (National Research Council, 2002, p. 22)

The report's primary outcome was the development of six guiding principles that "underlie all scientific inquiry, including education research" (National Research Council, 2002, p. 2). The report makes a case for the use of randomized field trials in education research under certain conditions arguing that experiments "are an ideal method when entities being examined can be randomly assigned to groups" (National Research

Council, 2002, p. 109). The authors' direct advocacy of randomized experiments can be found in this passage:

> In estimating the effects of programs, we urge the expanded use of random assignment. Randomized experiments are not perfect. Indeed, the merits of their use in education have been seriously questioned. For instance, they typically cannot test complex causal hypotheses, they may lack generalizability to other settings, and they can be expensive. However, we believe that these and other issues do not generate a compelling rationale against their use in education research and that issues related to ethical concerns, political obstacles, and other potential barriers often can be resolved. (National Research Council, 2002, p. 125)

Advocacy for the increased use of randomized experiments in education research and arguments for the preeminence of randomized experiments among alternative research methods have drawn the greatest criticism in these debates. Numerous scholars have countered the preeminence argument, advocating for a plurality of methods for understanding educational phenomena and contending that nonexperimental methods, particularly qualitative strategies, can provide direct evidence of causal effects (English, 2007; Gersten, 2013; Maxwell, 2004; Riehl, 2007). Similar arguments have been made regarding research on education leadership. Drawing on Borko's (2004) "situative" perspective, Riehl (2007) argues that understanding school leadership "requires a careful and contingent look at the circumstances and the enactment of leadership" (p. 146). Riehl and others have thus argued that this kind of nuanced, situated "look" at leadership within complex contexts is not possible with *pure* randomized control experiments.

Some critics of the use of randomized experiments in education research acknowledge the value of the method, but they do so with significant reservations. In summarizing his strong critique Maxwell (2004) writes, "None of this should be taken as denying or disparaging the value of experimental research in education" (p. 9). However, Maxwell (2004) goes on to describe three requirements for drawing causal inferences when using "strictly experimental designs, with no qualitative components" (p. 9):

> First, there should be a well-developed theory that informs the intervention and research design and allows interpretation of the experimental results. Second, the causal process investigated should be manipulable, fairly straightforward and simple and relatively free from temporal and contextual variability. . . . Third, the situation should not be conducive to the direct investigation of causal processes. (p. 9)

In clarifying what he means by "direct investigation" in the third requirement, Maxwell distinguishes between qualitative research procedures in which causal processes can be directly observed and research designs like experiments that depend on inferring causal relationships from the measured covariation of variables. Maxwell (2004) further argues that because these three conditions are often not met in education research, "in many instances, the optimal approach will be to combine qualitative and experimental methods" (p. 9), a position also advanced by the authors of *Scientific Research in Education* and in discussions of experiments in leadership research (Riehl, 2007).

We concur with many of the concerns raised about randomized experiments and attempted to address a number of them in designing the project reported here as a case study. Like some (English, 2007; Maxwell, 2004), we do not believe that randomized experiments hold a position of preeminence among research methods. Like others, however (e.g., Maxwell, 2004; Shadish, Cook, & Campbell, 2002), we believe that under certain conditions, experiments are particularly good at "causal description"—that is, causal inferences *that* an intervention had an effect. We further concur that pure experiments that do not include any "observational" variables and/or qualitative evidence beyond treatment group assignments are not well suited to providing causal explanations of *why* and *how* interventions have effects. The design of the project reported here as a case study included not only a randomized experiment but also observations of principals' daily work, qualitative interviews with principals, and survey measurements of the kinds of explanatory "causal processes" discussed by Maxwell (2004). We also collected evidence on the implementation of a professional development program for principals from interviews with district staff, interviews with national program staff, program attendance records, and program curricular materials. Evidence from these sources were used to illustrate themes from the literature just discussed. Thankfully, education research studies involving experiments increasingly include elements like these, thanks in part to the promotion of robust experimental designs by funders like the Institution of Education Sciences.

Experimental Evidence on Principals

An important part of the backdrop for this examination is the current state of experimental evidence on principals and principal development programs. While there has been increased advocacy of experiments in education research, their implementation in some subfields of inquiry has been limited,

and this is certainly true of the field of education leadership. To assess the availability of such evidence, we searched the Campbell Collaboration Social, Psychological, Educational & Criminological Trials Register (C2-SPECTR). C2-SPECTR contains abstracts of more than 10,000 randomized trials in the fields of sociology, psychology, education, and criminology. Abstracts contained in the C2-SPECTR database were identified by searching three major bibliographic databases (the Educational Research Information Clearinghouse, Sociological Abstracts, and Criminal Justice Abstracts) and 48 social science journals (Petrosino, Boruch, Rounding, McDonald, & Chalmers, 2000).

Our search of C2-SPECTR identified only three manuscripts that focus on principals in some manner. We searched C2-SPECTR using the terms *principal* and *leadership* and keywords *educational administration* and *educational supervision*. These searches identified a total of 18 articles. Of these, only three involved studies in which principals participated as subjects in a randomized experiment. One of these studies assessed principals' decision making in the teacher hiring process (Young, 1997). In this study, the background qualifications of fictitious teaching applicants were experimentally manipulated and principals' decisions were measured. The remaining two randomized trials both tested the effect of principals' participation in professional development on their practice. In an experiment reported in Thomas (1970), 28 principals were randomly assigned to participate in 5 days of training designed to improve their relationships with staff members. The study found that the group who attended the laboratory training altered their interpersonal behavior with their staff (Thomas, 1970). In the third randomized experiment, principals were randomly assigned to one of four conditions: (1) principals but not their teachers participated in a classroom management and a supervision workshop, (2) both teachers and principals participated in a classroom management workshop, (3) teachers participated in a classroom management workshop but principals did not, and (4) neither teachers nor principals participated in the workshop (Grimmett & Crehan, 1987). The results of this experiment indicated that supervision behavior was more effective when both teachers and principals participated in the workshop training.

Another randomized experiment involving principals is currently coming out of the field. For the *School Leadership Improvement Study*, principals were randomly assigned to either receive or not receive the Balanced Leadership program (Jacob, Goddard, Kim, Miller, & Goddard, 2014). Their study found that principals who participated in the program reported feeling more effective and using more effective practices than control group principals. Principals and teachers from participating

schools were also more likely to remain in the schools. However, there was no impact of the program on school instructional climate and student achievement. We note that a second major repository of experimental evidence on educational interventions, the What Works Clearinghouse, systematically excludes studies that do not examine student outcomes and, thus, by design does *not* include any studies that only examine the effects of professional development programs on principals or teachers. In addition to searching C2-SPECTR, a search of Google Scholar failed to identify any additional studies. We conclude from this literature search that experimental evidence on principals, their practice, and the effect of principal training programs (either professional development or preservice) is virtually nonexistent.

Common Challenges in Implementing Experiments in Education

Many aspects of an experiment can be compromised during implementation, but perhaps no aspect is more important than the randomization of subjects. The randomization process can be subverted at the outset of the study when subjects are randomly assigned to treatment conditions. Boruch (1997) reports on a medical study in which random assignment was left to hospital admissions staff. Staff members subverted random assignment by systematically assigning patients with more intense symptoms to the treatment group. The hospital staff members failed to implement the random assignment procedure because they believed that patients assigned to the control group would benefit from the experimental treatment. Boruch argues that the best way to avoid subversion of this sort is for researchers to exercise as much control over the random assignment process as possible.

A second common challenge in implementing experiments is subject compliance with treatment protocols. Bloom (2005) provides a classification scheme describing different forms of noncompliance. “Compliers” are subjects who conform to their random assignment either by receiving a treatment when assigned to a treatment group or by not receiving a treatment if they are assigned to a control group. “No-shows” are subjects who are assigned to receive a treatment but do not receive it. “Crossovers” are research subjects who are assigned to the control condition but who receive treatment.

A third common challenge in conducting experiments is ensuring both the quantity (dosage) and quality of treatment that is delivered. Any reduction in dosage to treatment group assignees or receipt of dosage by control group

assignees can greatly diminish a researcher's ability to detect treatment effects. Consequently, Boruch (1997) urges researchers to document treatment delivery as carefully as possible, thereby measuring the amount or dosage of treatment received. The *quality* of treatment delivery can also differ from the intended treatment (i.e., the treatment is delivered in a different manner from that called for by the treatment protocol). For example, Hulleman and Cordray (2009) found that the magnitude of the effects of an intervention designed to increase student motivation through writing tasks varied depending on the treatment dose (the number of opportunities students had to engage in the writing activities) and measures of the *quality* of students' engagement in the intervention.

A fourth common challenge with experiments conducted outside the laboratory is that treatment delivery is susceptible to changes in the social or policy context within which the experiment is conducted. Rossi, Lipsey, and Freeman (2004) describe the volatility of sociopolitical contexts this way:

> One of the most challenging aspects of program evaluation is the continually changing decision-making milieu of the social programs that are evaluated. The resources, priorities, and relative influence of the various sponsors and stakeholders of social programs are dynamic and frequently change with shifts in political context and social trends. (p. 22)

In the case of educational programs, the degree to which educators in schools embrace and implement such programs will greatly depend on the extent to which the school district endorses and supports the program. Rossi et al. (2004) go on to argue that evaluators should size up the social and political context before implementing an experiment to gauge the practicality of implementing an experimental evaluation in that context. They also urge researchers to be flexible and ready to adapt to changes in the social and political landscapes that are very likely to occur.

Case Study of an Experimental Evaluation of Principals

This case study outlines our mixed-methods approach to testing the effect of the professional development program on participants' leadership practice and presents the results of the experiment. We then illustrate how a range of factors affected program implementation and, subsequently, our ability to observe program effects.

The experimental evaluation of a District Professional Development program (DPD) was conducted during the 2004-2005, 2005-2006, and 2006-2007 school years. According to DPD's developers, the overarching goal of the program is to develop school leaders who will drive their schools to high performance through sustained improvements in instruction. Analysis of DPD curriculum materials indicated that the major themes of the intended curriculum include the principal as strategic thinker, principal as instructional leader, and principal as creator of a just culture in which all students achieve the same high standards. These themes are threaded throughout the program's 14 units.

The program incorporated many design principles that boded well for its potential impact on principals. Learning experiences were "situated," involving peer collaboration and ample opportunities for applying what was learned in context. Principals had opportunities to work on topics over long periods of time. The program also used face-to-face instruction in workshops, study groups, case studies, and action projects, as well as distance learning experiences. Educational experts were also featured prominently in the curriculum.

The DPD was created by a national developer. Typical program delivery involves staff from the district and DPD. In the first stage of delivery, national staff from the program train a local leadership team and provide technical assistance as that team subsequently trains cohorts of principals and other school leaders. Leadership teams typically include a project director, principals at each level (elementary, middle, high school), district or state administrators in curriculum and instruction, and sometimes local university faculty members. In addition to conducting the institutes for leadership teams, national program staff also provide substantial postinstitute assistance by facilitating training sessions as needed. Local training typically continues for 18 months to 2 years, although it may be compressed or expanded to fit local needs.

Sources of Evidence

We utilized five sources of evidence in developing the case study.

Interviews with district officials. Throughout the course of the study, we conducted semiregular interviews with key district staff members who were responsible for professional development in the district. These staff members included the chief academic officer, deputy superintendent for curriculum and instruction, and the director of professional development. These interviews captured the district's priorities for principal professional development

and how these priorities changed over time. The interviews also revealed the nature of the district's support for the DPD over the course of the study.

Review of DPD curriculum materials. We reviewed DPD curriculum materials provided to principals, including online learning modules. Our review of these materials was used to map out the *intended* DPD curriculum.

Observations of DPD training. DPD training sessions were observed, and observers wrote detailed narrative descriptions of the learning activities and content covered during each session. The description of the content covered at these sessions provided a record of the *delivered* DPD curriculum. Comparing the curriculum that was actually delivered to the intended curriculum provided a direct record of the quantity and quality of treatment dose.

DPD training attendance records. We also monitored treatment delivery by compiling attendance records for all DPD training sessions. These records documented principals who were assigned to receive the DPD treatment but failed to do so ("no-shows") and those principals who were not supposed to attend DPD trainings but did so ("crossovers"). Attendance records also captured the number of professional development sessions attended by each principal that we used as a measure of treatment dose.

Daily principal logs. All principals in the district were asked to complete web-based logs, which captured how they allocated their time across nine leadership domains for each hour of a school day. Principals completed daily logs during seven periods between spring 2005 and spring 2007, completing one log per day for five consecutive school days during each period. Multilevel analysis of data from the daily log was used to quantitatively estimate the effect of DPD on principals' leadership practice. Our use of the log instrument for this purpose is discussed in greater detail below.

An in-depth understanding of the DPD program's curriculum and intended delivery was developed through a review of DPD curriculum materials and interviews with DPD developers. A chronological narrative of implementation of DPD in the district under study was developed by summarizing DPD training observation notes, DPD attendance records, and transcripts of DPD district staff interviews. The creation of this narrative involved straightforward extraction, summarization, and synthesis of information and did not involve strong inferences or interpretation. By examining these complementary sources of evidence, we sought to develop a detailed narrative of DPD implementation that could be used to examine and illustrate benefits, limitations, and challenges identified by the literature.

Experimental Design

The experiment reported here is part of a larger mixed-methods longitudinal study that was conducted in Cloverville (a pseudonym), a midsized urban school district in the southeastern United States. In an attempt to reduce the chances that the experiment would be adversely affected by changes in the social or policy context, Cloverville was selected in part because the superintendent was a staunch supporter of the DPD and our experimental evaluation. Furthermore, when the experiment was approved by the district, the superintendent had 2 years left on a 3-year contract, so it was expected that support for the program and the randomized evaluation would continue for the duration of the study.

The original research plan was to use a delayed-treatment design in which half of Cloverville principals were randomly assigned to begin participating in the DPD at the outset of the study (early-treatment group), and half were randomly chosen to begin the DPD 1 year after the first group (delayed-treatment group). The design became a simple randomized experiment when the Cloverville district decided not to deliver the program to the second group of principals. The research design followed principals for three school years beginning in 2004-2005.

After excluding principals who were members of the Cloverville DPD leadership team, the remaining 48 principals in the district were randomly assigned to either the early-treatment or delayed-treatment groups. We used a basic random assignment design that incorporated school level (i.e., elementary, middle, high) as a blocking variable. To prevent the subversion of randomization, a member of the research team performed the random assignment. After random assignments were made, the randomization process was checked by comparing the two groups of principals on a wide range of variables measuring school and principal characteristics, including gender, race, years of experience, and whether the school had met adequate yearly progress. These comparisons demonstrated that principals assigned to the two groups were nearly identical on every variable examined.

We also instituted procedures for retaining sample members over the 3 years of the study. During each year of the experiment, we had multiple points of contact with Cloverville principals. Principal transfers both within and out of the district were carefully recorded. Boruch (1997) argues for the importance of providing financial incentives when participation is voluntary to retain sample members. The DPD study had multiple components, and principals received separate incentive payments for participating in each component. Principals who participated in all study components received incentive payments totaling $235 per year.

Table 1. Characteristics of the Schools of 48 Principals in the Experimental Sample.

Demographic Characteristic	Mean	Standard Deviation
School size	644	301
Percent Black	67	26
Percent Hispanic	3	4
Percent free/reduced price lunch	59	21

As can be seen in Table 1, even though all schools were located in the same urban district, there was substantial variation in their demographic characteristics. The average student enrollment for the schools of the 48 principals was 644, though the standard deviation of 301 indicated a substantial range across schools. On average, the schools of principals in the experimental sample had an African American enrollment of 67%, although the standard deviation of 26% indicates a broad range of student ethnicity in schools.

Challenge 1: The Effects of Policy Shifts

Even before it began, the experiment was affected by the kinds of policy shifts identified by Rossi et al. (2004). Initially, our proposed research was to be conducted in Brockville (a pseudonym), which is among the 20 largest school districts in the United States. At the time we proposed the DPD study, Brockville had already developed a leadership team and had begun piloting the DPD curriculum. But despite having a long relationship with the DPD developer, Brockville chose not to expand DPD delivery beyond the initial group, opting instead to use its own leadership development program.

On learning that Brockville was no longer offering DPD, we were able to successfully negotiate with the Cloverville district to serve as a site for the DPD evaluation. However, in conducting the study in Cloverville, our sample was reduced from 60 to 48 principals, thus considerably reducing statistical power. At the time, Cloverville was the largest of four districts across the country that had agreed to adopt the DPD program and was therefore the most suitable research site available. In addition, we had to locate a district that had not yet begun the DPD program so we could implement random assignment and the collection of pretreatment measures.

During the summer and fall of 2004, the Cloverville leadership team participated in institutes provided by DPD staff. Interviews with district

staff and observations of these preliminary trainings indicated that most of the leadership team was enthusiastic about DPD and that principals on the team began testing out some of the program's ideas in their schools during the fall of 2004. In the spring of 2005, the sample for the DPD experiment was selected and principals were randomly assigned by members of our research team to early-treatment and delayed-treatment groups as previously discussed. During the spring of 2005, we collected baseline measures on principals using daily logs, surveys, and observations, and in the summer of 2005, members of the leadership team participated in another institute provided by DPD staff.

In 2005, implementation of the DPD and the experimental evaluation continued to be affected by shifts in district priorities. In the summer of 2005, the Director of Professional Development informed us that the Cloverville superintendent was forced to resign. The district hired a new superintendent who began in the fall of 2005. The departing superintendent, Mr. Anderson (pseudonym) was largely responsible for bringing DPD to the district. Interviews with Mr. Anderson's Deputy Superintendent for Curriculum and Instruction revealed that Anderson had a long-standing relationship with DPD staff prior to coming to Cloverville, and he saw DPD as the primary vehicle for developing school leaders in the district.

Interviews with the executive director for professional development who stayed on in the district after Anderson's departure revealed that the new superintendent, Mr. Johnson (pseudonym), had no prior experience with the DPD program and brought his own ideas and preferences for leadership development to the district. The Director of Professional Development left soon after the new superintendent was hired to pursue an opportunity in a larger urban district. One major consequence of the arrival of a new superintendent was that district leadership decided that DPD training would *not* be provided to the second cohort of principals (i.e., the delayed-treatment group in the experiment). From the study's perspective, this meant that the design was no longer a delayed-treatment design but, instead, was a simple randomized experiment, since the early-treatment group was now the only group of principals who would receive treatment. Through subsequent interviews with principals and new district leaders, we learned that the new superintendent had his own professional development priorities and initiatives for school principals. He began to bring to Cloverville some of the training he had used in his previous district and showed little engagement with DPD.

Principals were originally intended to receive 2 years of DPD training. Conversations with district leadership in the spring of 2006 indicated that principals in the early-treatment group would be permitted to finish the full

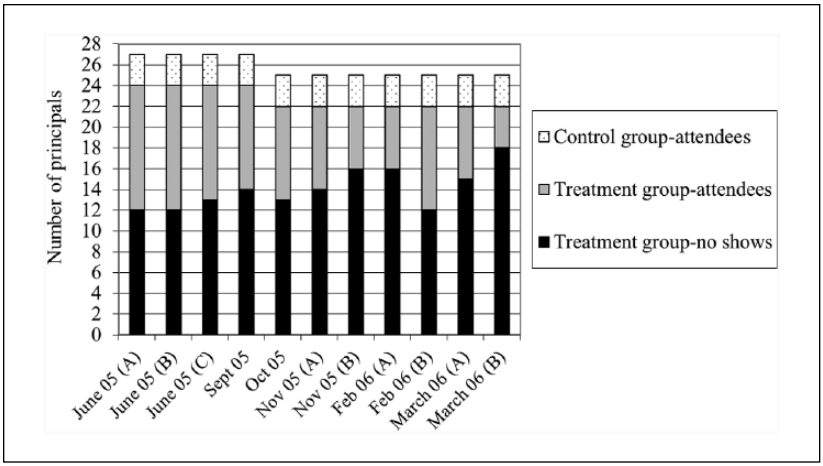

Figure 1. DPD attendance by treatment group and session.

18-month course of DPD training but that principals in the delayed-treatment group would not begin the program. Had this plan been enacted, early-treatment group principals would have continued to receive DPD training during the 2006-2007 school year. Despite these plans, no DPD training was provided during the 2006-2007 school year. Ultimately, only half of the sessions constituting the DPD curriculum were delivered.

Challenge 2: Subversion of Random Assignment

In June 2005, members of the leadership team delivered the first units of the DPD curriculum to principals (Figure 1 displays the attendance for each session). At the initial training, observers found evidence of the subversion of treatment assignment and nonparticipation. Of the 24 principals assigned to the early-treatment group, only 12 attended the first DPD training in June. This pattern continued as 10 of the initial 12 "no shows" did not attend a single DPD training. While only half of the principals who were expected to attend training did so, three principals who were *not* assigned to attend DPD training did so. During the 2005-2006 school year, an additional five units of the DPD curriculum was delivered by members of the leadership team. Similar patterns of attendance that were evident at the first session in June were observed in these subsequent training sessions. We were unable to determine whether the subversion of random assignment

occurred at the district level or whether these three principals simply sought out the DPD training on their own accord.

Challenge 3: Reduced Program Delivery

In total, 11 DPD training sessions were delivered to Cloverville principals. Attendance at DPD sessions decreased over time. The number of eligible principals in the treatment group also declined over time through a small number of retirements. Principals assigned to the treatment group had an overall attendance rate of about 42%.

As previously mentioned, the full DPD curriculum consists of 14 units. During the 2005-2006 school year, six curricular units and a unit on coaching were delivered. These delivered units constituted roughly half of the intended training sessions. Additionally, because of the spiraling nature of the curriculum, the omission of curricular units meant that participants did not receive the full intensity of topical coverage. Among the units that were delivered, there was a major emphasis on the use of strategic planning within a cycle of improvement to drive change in a school.

Results of the Experiment

We next turn to the results of the experiment assessing the effect of DPD on principal practice in Cloverville. Analyses compared principals assigned to treatment and control groups on nine domains of leadership responsibility at three time points. Drawing on DPD attendance records, interviews with district staff connected to the program, interviews with national DPD staff, we interpreted observed differences between the two groups within the context of changing district policies and DPD attendance patterns.

We examined the impact of the DPD on principals' practice in nine domains of leadership responsibility described in Table 2.

We view principal leadership practice as actions taken by principals to influence people, processes, and organizational structures, and we view the influence of principal leadership practice as being exercised through these multiple domains of responsibility. Like others, our measurement strategy attempted to take account of how principals distribute their time across a comprehensive set of domains (see, e.g., Drake & Roe, 2003; Martin & Willower, 1981; Peterson, 1977). Principal practice was measured with a daily web log both prior to and after the delivery of the DPD in the district. For each hour of the school day, the daily log captured the number of minutes principals spent on each of the nine domains of leadership. Measures of principals' emphasis in each domain were created by summing the number of

Table 2. Domains of Leadership Responsibility Measured by Daily Log Instrument.

Building operations (schedules, space allocation, building maintenance, vendors)
Finances and financial support for the school (preparing budgets, budget reports, seeking grants, managing contracts)
Community or parent relations (formal meetings and informal interactions)
School district functions
Student affairs
Personnel issues (recruiting, hiring, supervising, evaluating, problem solving)
Planning/setting goals (school improvement planning, developing goals)
Instructional leadership (monitoring or observing instruction, school restructuring or reform, supporting teachers' professional development, analyzing student data or student work, modeling instructional practices, teaching a class)
Your professional growth (formal professional development, attending classes at college/university, reading articles or books)

minutes principals spent in a domain each day a log was completed. Analyzing data from the same sample examined here, Camburn, Spillane, and Sebastian (2010) found that measurements of principal practice from the daily log closely matched evidence from an experience-sampling instrument and direct observations and concluded that the log instrument provided accurate measures of principals' work.

Based on an understanding of the content of the DPD curriculum, we originally hypothesized that on postbaseline measurements, principals assigned to the early treatment group would have increased the amount of time they devoted to instructional leadership and planning/goal setting. We also hypothesized that treatment group principals would have significantly higher postbaseline means on measures of instructional leadership and planning/goal setting than control group principals. We further hypothesized that the amount of time treatment group principals spent on other leadership functions that were not a primary focus of the DPD curriculum, such as finances, personnel, building operations, and student affairs, would not increase after their participation in DPD. However, the truncated delivery of the DPD curriculum meant that these hypotheses were no longer sensible. Observations of DPD training sessions indicated that among the units that were delivered, there was a major emphasis on the use of strategic planning within a cycle of improvement to drive change in a school. Key ideas presented in this training included the cycle of continuous improvement through strategic planning, including plan revision and the implementation of revised plans, and the monitoring of key outcomes, including instruction. Ultimately, a relatively small fraction of the program's content on instructional leadership was

delivered. Consequently, in conducting the analyses of the experiment, we no longer expected to observe positive effects of the program on principals' emphasis on instructional leadership. We did anticipate, however, that principals in the treatment group might exhibit increases on postbaseline measures of planning/goal setting and might spend more time on this leadership domain postbaseline than control group principals.

Analysis Model

We used data from the daily log to estimate two kinds of treatment effects. The first of these is referred to as the intent to treat (ITT). The most fundamental way to assess the effectiveness of a program using experimental data is to simply compare the means of treatment and control groups on a dependent variable that is believed to be affected by program participation. This comparison, which is based on subjects' *assignment* to treatment, regardless of actual *participation*, reflects the *intended* treatment plan, often referred to as the *intent to treat.* Boruch (1997) argues that there are both scientific and policy rationales for always conducting an ITT analysis, regardless of whether the treatment was delivered as intended and regardless of whether research subjects participated in the treatment as intended. The scientific rationale is that comparisons of randomly created groups will produce unbiased estimates of treatment impact that allow for legitimate causal statements about treatment effects. ITT analyses also have policy relevance because they reflect the kinds of imperfect treatment delivery and treatment participation that are likely to occur in a "real world" implementation of the program. In other words, an ITT analysis answers the policy question, "What is the likely effect of this treatment if it was *made available* to people like those studied?"

Alternatively, the second effect estimate we produce involves a simple comparison of people who participated in the DPD to those who did not. We operationalized participation as attending at least one DPD training versus attending none of the training sessions. Because this analysis ignores random assignment, it is not a causal estimate and the differences in outcomes between participants and nonparticipants may be confounded with unobserved variables.[1] However, this comparison is relevant in the present study given that researchers will often be interested in knowing how ITT estimates might be similar or different to comparisons between those who do and those who do not participate in professional development programs. This will be especially true when experiments suffer from noncompliance with treatment assignments.

Since data from the daily log have a nested structure of multiple daily reports per principal, we used a two-level model to estimate treatment effects on the principal leadership outcome measures. We employed the commonly used strategy of statistically controlling for a variable that is correlated with the outcome (in our case a pretreatment measure of the outcome) to increase statistical power for the impact estimate. Such controls increase statistical power by reducing the standard error of the impact estimates (Bloom, Bos, & Lee, 1999; Raudenbush, 1997). Statistical models also controlled for the blocking variable school level (elementary schools vs. middle schools and high schools). With ITT analyses, a dummy variable indicating treatment status was coded 1 if principals were assigned to the treatment group and 0 if they were assigned to the control group. For the participant/nonparticipant analyses, the treatment indicator variable was coded 1 if principals attended at least 1 DPD session and 0 if they did not attend any DPD sessions. The statistical equations for the multilevel models are included in the appendix.

Results

We first investigated the effect of the DPD on principals who were randomly assigned to participate in the program (the effect of the ITT). Table 3 displays the average minutes per day principals assigned to treatment and control groups spent on the nine leadership domains. For each domain, statistical models estimated the mean for principals assigned to the control group (column labeled "Control Group Mean" in Table 3) and the average difference between control group and treatment group principals (column labeled "Treatment Group Coefficient" in Table 3). Averages for principals assigned to the treatment group for each of the leadership domains were calculated by simply adding the difference between treatment and control groups to the mean for the control group ("Treatment Group Mean" in Table 3). Averages are displayed for two time points, spring of 2005 and spring of 2006. Since the delivery of the DPD program to principals began during summer 2005, means for spring 2005 are baseline pretreatment measurements and the spring 2006 measurements are posttreatment. The hypothesis tests reported in Tables 3 and 4 were not adjusted to account for the multiple comparisons that were made. In conducting multiple statistical tests using the same data set, the probability of finding statistical significant results purely due to chance increases with the number of hypothesis tests conducted. It is thus important to note that none of the hypothesis tests reported in Tables 3 and 4 would be significant after adjusting for multiple comparisons.

Table 3. Predicted Number of Minutes per Day Spent on Leadership Domains by Experimental Treatment Assignment (Intent to Treat).

	Spring 2005 Pretreatment					Spring 2006				
	N	Intercept Control Group Estimate	Standard Error	Treatment Group Coefficient	Standard Error	*N*	Intercept Control Group Estimate	Standard Error	Treatment Group Coefficient	Standard Error
Building operations	46	44.28	6.70***	−2.10	7.18	31	21.91	11.34	9.93	11.45
Community or parent relations	46	38.13	10.32***	12.15	11.05	31	29.67	10.05**	12.81	10.23
School district functions	46	27.42	6.21***	−13.42	6.65	31	32.64	15.24*	8.97	14.81
Finances	46	24.09	6.42***	6.89	6.88	31	13.51	5.53*	6.41	5.64
Instructional leadership	46	89.09	18.38***	−18.05	19.67	31	75.46	16.10***	7.48	16.16
Personnel issues	46	41.79	11.81***	9.96	12.65	31	40.59	17.52*	−25.46	18.26
Planning/setting goals	46	37.33	6.71***	2.81	7.18	31	54.09	24.10*	−5.28	23.86
Your professional growth	46	20.14	8.42 *	1.41	9.01	31	4.34	8.51	9.39	8.65
Student affairs	46	105.56	18.37***	3.75	19.67	31	126.15	16.66***	20.89	18.72

*$p < .05$. **$p < .01$. ***$p < .001$.

Table 4. Predicted Number of Minutes per Day Spent on Leadership Domains by District Professional Development Program Participation.

	Spring 2005 Pretreatment					Spring 2006				
	N	Intercept Participant Estimate	Standard Error	Participant Coefficient	Standard Error	*N*	Intercept Participant Estimate	Standard Error	Participant Coefficient	Standard Error
Building operations	46	44.56	6.45***	−3.29	7.52	31	20.03	11.18	14.72	11.38
Community or parent relations	46	45.58	10.08***	−4.19	11.78	31	38.07	10.27***	−8.46	10.56
School district functions	46	22.58	6.24***	−4.00	7.27	31	23.46	14.59	31.12	14.17*
Finances	46	28.59	6.24***	−3.11	7.30	31	15.90	5.74*	−0.28	5.78
Instructional leadership	46	73.61	17.60***	17.36	20.60	31	73.76	15.65***	14.74	17.03
Personnel issues	46	48.67	11.45***	−5.49	13.38	31	39.03	17.57*	−22.62	18.98
Planning/setting goals	46	44.12	6.25***	−14.04	7.27	31	26.69	23.09	55.79	22.79*
Your professional growth	46	18.13	8.09*	6.92	9.41	31	9.74	8.71	−4.94	8.95
Student affairs	46	112.19	17.57***	−12.40	20.56	31	133.20	17.32***	−5.38	17.74

*$p < .05$. **$p < .01$. ***$p < .001$.

We found that in the spring of 2005, none of the difference in the number of minutes treatment and control group principals spent on the nine leadership domains were statistically significant. As previously reported, principals assigned to treatment and control groups were also nearly identical on a large number of principal and school characteristics.

In the spring of 2006, treatment group principals spent more time than control group principals on a number of domains that were not a major focus of the DPD curriculum. In particular, they were estimated to spend 10 minutes more per day on building operations and 12 minutes more on community or parent relations. They were also predicted to spend 25 minutes less on personnel issues. While the size of these differences struck us as practically significant, none were statistically significant. Principals in the two groups spent similar amounts of time on most of the other leadership domains in spring 2006. Given that participation in the DPD required considerable time of principals, we expected to see principals assigned to participate in the DPD spending more time on their personal professional growth. We did in fact observe this pattern. In spring 2006, principals assigned to the treatment group were estimated to spend more time on their own professional growth than principals assigned to the control group, though these differences were not statistically significant.

We were particularly interested in whether treatment group principals spent more time on instructional leadership or planning/goal setting during the year in which the DPD was implemented in Cloverville. Recall that these two leadership domains were major foci of the DPD curriculum. We found that principals assigned to the two groups devoted roughly the same amount of time to instructional leadership in spring 2006. Principals assigned to participate in the DPD were actually predicted to spend *less* time on planning/setting goals than those assigned to the control group, despite the major emphasis on this aspect of leadership in the portion of the DPD program that was actually delivered.

We next compared the leadership practice of principals who participated in the DPD program with those who did not participate. The results of these analyses are displayed in Table 4 following the same format used for Table 3. We remind readers that the results in Table 4 are not causal estimates of program effects and are thus simply descriptive results about the association between program participation and principal leadership practices.

In spring 2006, DPD participants were estimated to spend 30 minutes more per day working with Cloverville district staff members. We believe that these differences may be partly related to the intervention. Recall that the DPD was delivered by a leadership team of which district staffs were members. Also, many of the DPD sessions were held at the central district office.

Table 5. Descriptive Analysis of Principal Characteristics Associated With Treatment Compliance.

	Treatment Compliers			Control Compliers			No-Shows			Crossovers		
	N	Mean	*SD*	*N*	Mean	*SD*	*N*	Mean	*SD*	*N*	Mean	*SD*
Leadership experience	13	10.85	5.29	17	13.88	6.59	7	18.71	4.46	3	8.67	1.16
School needs improvement	13	0.23	0.44	20	0.35	0.49	8	0.38	0.52	3	0.33	0.58
Efficacy	12	13.09	1.24	19	12.95	1.68	7	11.00	2.89	3	13.33	2.89
Regard/value for district changes	12	5.92	1.16	18	5.89	1.32	7	4.86	1.68	3	5.33	1.15

Thus, Cloverville district staff played a central role in the delivery of the truncated curriculum. Surprisingly, the amount of time principals who participated in the DPD program spent on their professional growth did not differ substantially from their nonparticipating counterparts.

We also found that in the spring of 2006, principals who participated in the DPD spent considerably more time planning and setting goals than their nonparticipating colleagues. Nonparticipating DPD principals spent 26 minutes per day in spring 2006, whereas participating principals were predicted to spend approximately 82 minutes per day on this leadership function. The fact that the largest absolute difference between participating and nonparticipating principals was observed in the leadership domain receiving the greatest emphasis in the truncated curriculum after participating principals had received the maximum amount of training is suggestive that participating principals might have been affected by their participation in the DPD.

Earlier we discussed how diminished treatment delivery likely reduced the impact of DPD. We were also interested in how the program's impact might have been influenced by the complex pattern of DPD *participation* in Cloverville. Analysis of DPD attendance records and qualitative interviews with principals suggested that DPD attendance might have been influenced by an interrelated set of factors—principals' career stage, the value they placed on district initiatives, and their incentives and motivation to improve their practice. We developed four proxy measures[2] of these factors and used descriptive statistics to examine how treatment compliance varied by these variables (Table 5).

We found that principals who were assigned to treatment but did not attend (labeled "No-Shows" in Table 5) had considerably more leadership seniority than those in the other groups. Principals in this group also were more likely

to be leading schools designated as being "in need of improvement," had lower levels of efficacy, and had much lower opinions of the districts' improvement efforts than any other group. In contrast to these nonattendees, principals who were randomly assigned to participate in the DPD program and actually attended (labeled "Treatment Compliers" in Table 5) were less senior, had a higher sense of self-efficacy, and had more positive opinions of Cloverville's initiatives. The three principals who were assigned to the control group but who attended DPD training anyway (labeled "Crossovers" in Table 5) more closely resembled "Treatment Compliers" than "No-Shows, "being generally more junior, more efficacious, and holding district initiatives in higher regard than the latter group. While this descriptive evidence is far from conclusive, these results provide a potential window into the intervention selection process and suggest that more junior, highly motivated principals in less challenging settings may have found the program more attractive and less motivated, veteran principals in more challenging settings may have found the program less attractive.

Discussion

There is widespread agreement that the primary *benefit* of experiments in education is that they provide a strong basis for deciding whether or not a program has achieved its intended result. More specifically, ITT analyses conducted for experiments provide beneficial policy guidance by answering the question "What is the likely effect of this treatment if it was *made available* to people like those studied?" The answer to this question provided by the experimental evaluation of DPD is that the program would likely have *no* effect on principals' emphasis on instructional leadership or planning. Additional analyses suggested that the DPD may have had a short-term impact on the amount of time principals spent planning and setting goals. However, this effect was only significant in the noncausal analyses comparing participants and nonparticipants, and disappeared after adjustments for multiple comparisons were made.

One of the most significant *limitations* of pure experiments is that they do not provide a solid causal basis for deciding *how* programs work. Weiss (1995) argues that evaluations are useful when they "show which of the assumptions underlying the program break down, where they break down, and which of the several theories underlying the program are best supported by the evidence" (p. 67). Capturing this kind of evidence and documenting program implementation is difficult. The potential pathways through which leadership development initiatives might influence principals, teachers, and students are extraordinarily complex and typically involve (1) multiple program components, (2) multifaceted

meanings and definitions, (3) complex psychological and social processes, and (4) both individual and organizational change (Gutierrez & Tasse, 2007). Given such complexity, experimental designs that merely provide a yes/no answer about a program's effectiveness will be of limited utility to policymakers, leaders, and practitioners. Augmenting experimental designs with evidence documenting how programs are implemented has the potential to overcome this limitation (Goldring, Preston, & Huff, 2013). This kind of augmentation can be considered a form of data triangulation. Advocates of mixed-methods research designs argue that all methods have limitations and biases, and many regard triangulation as a useful way to increase the validity of research results by minimizing these limitations and biases (Denzin, 1989; Webb, Campbell, Schwartz, & Sechrest, 1966). Our augmentation of experimental evidence from the DPD evaluation with evidence from DPD staff interviews, attendance records, and curricular materials permitted us to go beyond the experimental results and begin to develop an understanding of *why* the DPD failed to have an effect. Similarly, using data on the same sample of principals, Barnes, Camburn, Sanders, and Sebastian (2010) went beyond experimental evidence by using data on principals' implementation of DPD from interviews and observations to understand *how* the DPD influenced principals. In doing so, they documented how principals used knowledge structures, tools, and routines provided by the DPD to make modest refinements in their practice.

There are also substantial limitations in how broadly applicable the results of experiments are. Strictly speaking, experimental results are only generalizable to members of the treatment group and to the set of conditions in which the treatment was delivered, received, and implemented. These limitations have very practical implications. Decision makers need to know how programs work, and they need to have some degree of certainty about whether a program will work with the personnel who will implement the program in the prevailing conditions in their local setting. The experimental evaluation of DPD was limited in each of these ways. The results of the DPD experiment are only applicable to districts similar to Cloverville that implement the DPD with a similar group of principals in the same way principals in Cloverville implemented the program. Experiments like ours that are conducted in a single district face a further challenge of being at much greater risk of contamination where information about the intervention is shared with members of the control group. The exposure of the three "crossover" principals to DPD are an example of such contamination.

While experimental designs can be more elegant and straightforward to execute than other study designs, and less burdensome on study participants, in practice, researchers who conduct experiments face a multitude of *challenges*. Here we highlight five challenges we believe are particularly

consequential and reflect on our experiences with each challenge in conducting the experimental evaluation of DPD.

First, truncated treatment delivery and lack of treatment compliance pose major challenges to interventions and experimental evaluations of interventions. Clearly the biggest factor undermining our ability to detect an effect of the DPD program on principals was that only half of the intended program was delivered, and it was delivered to a smaller number of principals than intended. Our ability to detect an effect of the DPD program was further undermined by the fact that a number of principals who were not assigned to the treatment participated in the training, which diluted the power of intent-to-treat analyses. The difficulties we experienced are not unique. Noncompliance in experimental studies is quite common—for example, as many as 30% of patients in medical experimental studies have been found to fail to follow treatment assignment (Armitage, 1983; Cockburn, Gibberd, Reid, & Sanson-Fisher, 1987). At the same time, noncompliance and the factors underlying noncompliance can themselves be of interest to researchers as this information can be utilized for the design and planning of future experimental studies. Our descriptive analyses suggest that there may have been reluctance among more senior, less motivated principals in Cloverville to engage in the major changes in leadership practice required by the DPD program. To be broadly successful, professional development programs like DPD need to engage principals at all career stages and at all motivation levels. We believe the DPD attendance data suggest that further research is needed that investigates ways of inducing principals' participation in professional development.

Failure of program participants to implement programs with fidelity is a second major challenge facing experimental evaluations of principal professional development programs. Fidelity of implementation refers to how well the enacted intervention achieves the essential "core components" required to make the program work (Nelson, Cordray, Hulleman, Darrow, & Sommer, 2012; Wallace, Blase, Fixsen, & Naoom, 2008). The presence of these core components needs to be measured for both treatment and control groups. As Nelson et al. (2012) note,

> In order to make the causal claim that the presence of the intervention causes differences in outcomes, the extent to which intervention core components exist within the control condition also needs to be assessed. As with most educational and behavioral interventions, the core components are unlikely to be completely novel and may overlap with best practices or business as usual. Thus, core components of the intervention may appear to some degree in the control condition, resulting in less contrast between conditions and possibly weaker effects. (p. 377)

Cloverville's failure to deliver the full complement of 14 curriculum units is a clear indication that the program was not implemented with fidelity.

A third major challenge to the fidelity of program implementation occurs when educational interventions are inevitably adapted by program enactors (Berman, 1981; Cohen, 1990). Berman (1981) argues that program adaptation and its implications for implementation fidelity depend on the nature of the innovation. Echoing this idea, Cohen (2011) attributes the success in the implementation of comprehensive school reforms to designs that provided strong guidance and resources for teachers. Like many programs, the DPD program relies on local adaptation. The program is designed to impact teaching and learning primarily by promoting changes in principals' abilities to provide instructional leadership. The DPD design leaves much discretion for principals to promote improvements in teaching and learning in ways that best fit the needs of their school. In other words, by design, implementation of the DPD involves considerable local adaptation that leaves considerable ambiguity in what constitutes fidelity of implementation. When local adaptation is an important aspect of implementing an intervention, not only does assessing the fidelity of implementation become more difficult, understanding *how* an intervention influences desired outcomes becomes much more salient.

Sociopolitical conditions pose a fourth significant challenge to the implementation of interventions and experiments. In conducting the experimental evaluation of DPD, we were confronted with questions about the numerous challenges researchers face in evaluating large-scale educational programs in complex, regularly changing settings. Newmann, Smith, Allensworth, and Bryk (2001) describe a pervasive pattern whereby schools adopt a large number of incoherently related programs in a kind of "revolving door" fashion. This approach, which Rick Hess and others have called "policy churn," is a perpetual process of changing priorities and strategies at the district level (Hess, 1998). As we reported above, the experimental evaluation of DPD was adversely affected by shifting district priorities at a number of points. Experiments can sometimes buffer themselves against shifting priorities in the policy context by seeking out contexts that are supportive of the treatment (Borman, Slavin, & Cheung, 2005; Rossi et al., 2004). However, our experience in conducting the DPD experiment suggests that scouting out amenable contexts is not foolproof. As discussed above, Brockville and Cloverville were both very supportive of DPD when we initially approached these districts. However, in both settings, support for DPD eroded. And while our focus here has mainly been on the results of an experiment, we believe our experiences also highlight how district leadership changes can prevent, erase, or severely undermine the progress of policies and programs. Cloverville made substantial financial and human resource investments in the DPD

program, all of which were completely for naught when the new district superintendent took over. Of course, the losses to the district were very likely not completely financial, as the morale of principals who had invested considerable time in improving their practice through the DPD program was likely diminished when the program was cut short. A related challenge that was highlighted by experience is that the long time periods needed for most interventions to develop and exercise their influence increase the vulnerability of programs and program evaluations to the sociopolitical challenges just discussed.

Though we have limited direct evidence, our experience also suggests a fifth challenge—that principals are a particularly challenging population to study with experimental designs. As the chief executive officers of their schools, principals are busy, autonomous leaders of complex organizations. Programs like DPD, and studies that examine their effects, depend on the voluntary participation of these professionals who have many competing demands on their time. We were unable to determine the extent to which low DPD attendance rates among principals assigned to the treatment group were driven by constraints on principals' time, lack of interest in the DPD curriculum, or encouragement or discouragement from district staff. Future experiments targeting principals would be well advised to explore ways to encourage participation in programs to which principals are randomly assigned and to seek out research sites that are able to buffer programs from policy shifts for a suitable length of time so that more robust assessments of program effects can be performed.

Using regular contacts from field staff and considerable incentive payments, we experienced minimal attrition in the sample, only failing to retain three of 48 principals between spring 2005 and spring 2006. But retaining sample members is a minimal requirement in longitudinal experiments involving principals. Researchers also need to minimize nonresponse among retained sample members. In the spring of 2005, 93% of the principals in the sample completed daily logs, but in the spring of 2006, that number dropped to 70% despite a comprehensive regimen of phone calls, e-mails, and incentive payments. Among those who did not complete daily logs in spring 2006, most were simply too busy. Despite repeated contacts, we were unable to reach six principals. There were also eight other principals who told us they would complete their daily logs, but who ultimately could not find time to participate. In light of this experience, we believe that researchers conducting longitudinal experiments with principals would be wise to design data collection experiences that make study participation as convenient as possible. Even though observations and interviewer-administered instruments are more expensive than more commonly used self-administered instruments,

these data collection strategies can be designed to be less burdensome. Increasing investments in data collection budgets are well worth considering in our view if they lead to more complete and better quality data.

We would not have learned as much about the effectiveness of the DPD if the study's design had only included a randomized experiment. As it is intended to do, the experimental component of the design allowed us to form a trustworthy summary inference about *whether or not* the DPD had an overall effect on principals' practice in Cloverville. However, the randomized experiment component of our research design did not illuminate *why* or *how* the DPD failed to influence principal practice in the manner intended by district leaders and program developers. Using complementary data on the intended DPD curriculum, district staff interviews, and DPD attendance records, we were able to develop an understanding of why analyses from the experiment failed to detect hypothesized program effects and, more important, an understanding of why the professional development program failed to achieve its intended goals. In our view, the experiences we have detailed here support continued advocacy of research designs that bring rich evidence to bear about causal mechanisms, implementation conditions, potential measures of delivery of and adherence to treatment protocols, and measures of participants' exposure to treatment. We are encouraged by the increased promotion and use of such designs in education research. In our view, these kinds of enriched research designs will more effectively advance the field's understanding of how to design and implement professional development initiatives for principals and extend our understanding of how such initiatives work.

Appendix

Statistical Models

The equations used to estimate the hierarchical linear modeling models are as follows.

Level 1. In Level 1, the time spent on a particular domain (Building Operations, Finances, Student Affair, etc.) is the model outcome; there are no predictors at this level. In effect, this model estimates a weighted average for the number of minutes spent by a principal on a particular leadership domain across multiple work days.

$$(\text{Time on Leadership Domain})_{ij} = \beta_{0j} + r_{ij}$$

Level 2. At Level 2, β_{0j}, the average time spent by a principal on a particular leadership domain is predicted by (1) the average time spent on the same leadership domain prior to treatment (in spring 2005) and (2) group membership. Group membership depends on whether the analysis was an Intent to Treat or Participant/Nonparticipant analysis.

$$\beta_{0j} = \gamma_{00} + \gamma_{01} * (\text{Time on Leadership Domain in spring 2005}\, j) + \gamma_{02} * (\text{Group}_j) + u_{0j}$$

These analyses were carried out separately for all nine leadership domains. The prior measure of time spent on each leadership domain were standardized so that the intercept represented the time spent on a leadership domain by a principal who spent an average amount of time on that same domain in spring 2005 and who was either not assigned to treatment (for ITT analysis) or not treated by DPD (for Participant/Nonparticipant)

Declaration of Conflicting Interests

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

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was funded though a grant from the Institute of Education Sciences, award number R305E040085.

Notes

1. We do not present a complier average causal effect (CACE) estimates (Angrist, Imbens, & Rubin, 1996), which is commonly used in experiments that have no-shows and crossovers. This is because the CACE is equal to ITT/P_{CL}, where the denominator is the proportion of compliers; and therefore, nonsignificant ITTs will yield nonsignificant CACE effects. Thus, in this study, the CACE estimates provide little more information than the ITT estimates.
2. *Leadership experience* is the number of years the principal had worked as an administrator. *School needs improvement* is a dummy variable coded 1 if school was placed in formal improvement status by the state, Title I, No Child Left Behind, or the school district. *Efficacy* is the mean of four items measuring how much principals value changes required by their professional development and how capable they feel to make those changes. *Regard/value for district changes* is a combination of two items that capture principals' ratings of district initiatives.

References

The American Association of Colleges for Teacher Education. (2001, March). *PK-12 educational leadership and administration* (White paper). Washington, DC: Author.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*, 444-455.

Armitage, P. (1983). *Statistical methods in medical research*. Oxford, England: Blackwell Scientific.

Augustine, C. H., Gonzalez, G., Schuyler Ikemoto, G., Russell, J., Zellman, G. L., Constant, L., . . . Dembosky, J. W. (2009). *Improving school leadership: The promise of cohesive leadership systems*. Santa Monica, CA: RAND Corporation.

Barnes, C. A., Camburn, E. M., Sanders, B., & Sebastian, J. (2010). Developing instructional leaders: Using mixed methods to explore the black box of planned change in principals' professional practice. *Educational Administration Quarterly*, *46*, 241-279.

Berman, P. (1981). Educational change: An implementation paradigm. In R. Lehming & M. Kane (Eds.), *Improving schools: Using what we know* (pp. 253-286). London, England: Sage.

Bloom, H. S. (2005). *Learning more from social experiments: Evolving analytical approaches*. New York, NY: Russell Sage Foundation.

Bloom, H. S., Bos, J. M., & Lee, S.-W. (1999). Using cluster random assignment to measure program impacts: Statistical implications for the evaluation of education programs. *Evaluation Review*, *23*, 445-469.

Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. *Educational Researcher*, *33*(8), 3-15.

Borman, G. D., Slavin, R. E., & Cheung, A. (2005). Success for all: First-year results from the national randomized field trial. *Educational Evaluation and Policy Analysis*, *27*(1), 1-11.

Boruch, R. (1997). *Randomized experiments for planning and evaluation: A practical guide*. Thousand Oaks, CA: Sage.

Boruch, R. (2002). The virtues of randomness. *Education Next*, *2*(3), 36-42.

Camburn, E. M., Spillane, J., & Sebastian, J. (2010). Assessing the utility of a daily log for evaluations involving school principals. *Educational Administration Quarterly*, *46*, 707-737.

Cockburn, J., Gibberd, R. W., Reid, A. L., & Sanson-Fisher, R. W. (1987). Determinants of non-compliance with short term antibiotic regimens. *British Medical Journal*, *295*, 814-818.

Cohen, D. K. (1990). A revolution in one classroom: The case of Mrs. Oublier. *Educational Evaluation and Policy Analysis*, *12*, 327-345.

Cohen, D. K. (2011). *Teaching and its predicaments*. Cambridge, MA: Harvard University Press.

Condliffe Langeman, E. (2000). *An elusive science: The troubling history of education research*. Chicago, IL: University of Chicago Press.

Cook, T. (2002). Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis*, *24*, 175-199.

Crow, G. M. (2007). Foreword. In F. W. English & G. C. Furman (Eds.), *Research and educational leadership: Navigating the new National Research Council Guidelines* (pp. vii-x). Blue Ridge Summit, PA: Rowman & Littlefield Education.

Davis, S., Darling-Hammond, L., LaPointe, M. A., & Meyerson, D. (2005). *School leadership study: Preparing successful principals* (Review of Research). Stanford, CA: Stanford University, Stanford Educational Leadership Institute.

Denzin, N. K. (1989). *The research act: A theoretical introduction to sociological methods* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Drake, T. L., & Roe, W. H. (2003). *The principalship*. New York, NY: Macmillan College.

Eisenhart, M., & Towne, L. (2003). Contestation and change in national policy on "scientifically based" education research. *Educational Researcher*, *32*(7), 31-38.

Elmore, R. F. (2000, Winter). *Building a new structure for school leadership*. Washington, DC: Albert Shanker Institute.

English, F. W. (2007). Introduction. In F. W. English & G. C. Furman (Eds.), *Research and educational leadership: Navigating the new National Research Council Guidelines* (pp. xi-xv). Blue Ridge Summit, PA: Rowman & Littlefield Education.

Gersten, R. (2013). From the editor in chief: The two cultures of educational research? *Elementary School Journal*, *114*, 139-141.

Goldring, E., Preston, C., & Huff, J. (2013). Conceptualizing and evaluating professional development for school leaders. *Planning and Change*, *43*(3), 1-13.

Grimmett, P. P., & Crehan, E. P. (1987, May). *Changes in elementary school principals as a result of laboratory training*. Paper presented at the annual meeting of the Canadian Association for Teacher Education, Hamilton, Ontario, Canada.

Gutierrez, M., & Tasse, T. (2007). Leading with theory: Using a theory of change approach for leadership development evaluations. In K. M. Hannum, J. W. Martineau, & C. Reinelt (Eds.), *The handbook of leadership development evaluation* (pp. 48-70). San Francisco, CA: Jossey-Bass.

Hess, F. M. (1998). *Spinning wheels: The politics of urban school reform*. Washington, DC: Brookings Institution Press.

Howe, K. R. (2005). The question of education science: Experimentism versus experimentalism. *Educational Theory*, *55*, 307-321.

Hulleman, C. S., & Cordray, D. S. (2009). Moving from the lab to the field: The role of fidelity and achieved relative intervention strength. *Journal of Research on Educational Effectiveness*, *2*, 88-110.

Jacob, R., Goddard, R., Kim, M., Miller, R., & Goddard, Y. (2014). Exploring the causal impact of the McREL Balanced Leadership Program on leadership, principal efficacy, instructional climate, educator turnover, and student achievement. *Educational Evaluation and Policy Analysis*. Advance online publication. doi:10.3102/0162373714549620

LaPointe, M., Meyerson, D., & Darling-Hammond, L. (2006, April). *Preparing and supporting principals for effective leadership: Early findings from Stanford's School Leadership Study*. Paper presented at the 2006 annual meeting of the American Educational Research, San Francisco, CA.

Levine, A. (2005). *Educating school leaders*. New York, NY: Education School Project.

Martin, W. J., & Willower, D. J. (1981). The managerial behavior of high school principals. *Education Administration Quarterly*, *17*, 69-90.

Maxwell, J. A. (2004). Causal explanation, qualitative research, and scientific inquiry in education. *Educational Researcher*, *33*(2), 3-11.

Nelson, M. C., Cordray, D. S., Hulleman, C. S., Darrow, C. L., & Sommer, E. C. (2012). A procedure for assessing intervention fidelity in experiments testing educational and behavioral interventions. *Journal of Behavioral Health Services & Research*, *39*, 374-396.

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

National Research Council. (2002). *Scientific research in education* (R. Shavelson & L. Towne, Eds.). Committee on Scientific Principles for Educational Research. Washington, DC: National Academy Press.

Peterson, K. D. (1977). The principal's tasks. *Administrator's Notebook*, *26*(8), 1-4.

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

Petrosino, A., Boruch, R. F., Rounding, C., McDonald, S., & Chalmers, I. (2000). The Campbell Collaboration Social, Psychological, Educational and Criminological Trials Register (C2-SPECTR) to facilitate the preparation and maintenance of systematic reviews of social and educational interventions. *Evaluation and Research in Education*, *14*, 206-219.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, *2*, 173-185.

Riehl, C. (2007). Research on educational leadership: Knowledge we need for the world we live in. In F. W. English & G. C. Furman (Eds.), *Research and educational leadership: Navigating the new National Research Council Guidelines* (pp. 133-168). Blue Ridge Summit, PA: Rowman & Littlefield Education.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A systematic approach*. Thousand Oaks, CA: Sage.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher*, *31*(7), 15-21.

Thomas, T. A. (1970). *Changes in elementary school principals as a result of laboratory training*. Eugene: University of Oregon, Center for Advanced Study of Educational Administration.

Wallace, F., Blase, K., Fixsen, D., & Naoom, S. (2008). *Implementing the findings of research: Bridging the gap between knowledge and practice*. Alexandria, VA: Educational Research Service.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (1966). *Unobtrusive measures*. Chicago, IL: Rand McNally.

Weiss, C. H. (1995). Nothing as practical as good theory: Exploring theory-based evaluation for comprehensive community initiatives for children and families. In J. I. Connell, A. C. Kubisch, L. B. Schorr, & C. H. Weiss (Eds.), *New approaches to evaluating community initiatives: Concepts, methods, and contexts* (pp. 65-92). Washington, DC: Aspen Institute.

Young, I. P. (1997). Holmes versus traditional teacher candidates: Labor market receptivity. *Journal of School Leadership*, *7*, 330-344.

Author Biographies

Eric M. Camburn is a professor in the Department of Educational Leadership and Policy Analysis at the University of Wisconsin–Madison and a project leader at the Wisconsin Center for Education Research. His research focuses on factors that support instructional improvement and survey research methods in education research.

Ellen Goldring is the Patricia and Rodes Hart Professor and Chair, Department of Leadership, Policy and Organizations, Peabody College, Vanderbilt University. Her research interests focus on the intersection of education policy and school improvement with particular emphases on education leadership.

James Sebastian is an assistant professor in educational leadership and policy analysis at the University of Missouri–Columbia. His research interests include school organization, organizational theory and behavior, and urban school reform.

Henry May is the director of the Center for Research in Education and Social Policy and associate professor at the University of Delaware. He specializes in large-scale randomized experiments and the application of modern statistical methods in studying the implementation and impacts of educational and social interventions and policies.

Jason Huff is the director of Leadership Development for Seattle Public Schools. His work focuses on training aspiring and novice principals as well as principal assessment and evaluation strategies to build their instructional leadership capacity.