

DRAFT Report– Please do not cite or distribute without permission of the authors

Lessons Learned from an Experimental Evaluation of a Principal Professional Development Program

Eric M. Camburn
Ellen Goldring
Henry May
Jonathan Supovitz
Carol Barnes
James P. Spillane

This paper was presented at the 2007 meeting of the American Educational Research Association. The research was conducted by the Consortium for Policy Research in Education (CPRE) supported through a grant from the U.S. Department of Education’s Institute of Education Sciences. The work was also supported by a grant from the National Science Foundation (Grant # EHR – 0412510). Opinions expressed in this paper are those of the authors and do not necessarily reflect the views of districts, schools, or subjects studied; CPRE; or its institutional members. This paper was internally reviewed by CPRE research staff; however, it has not been externally reviewed.

We wish to thank Jason Huff, Beth Sanders, and Jimmy Sebastian for their research assistance at various stages of the work. Please direct any correspondence regarding this paper to Eric Camburn, camburn@wisc.edu. Eric Camburn is at the University of Wisconsin-Madison, Ellen Goldring is at Vanderbilt University, Henry May and Jon Supovitz are at the University of Pennsylvania, Carol Barnes is at the University of Michigan, and James Spillane is at Northwestern University.

In recent years critics have argued that leadership preparation in this country often fails to prepare future school leaders for the modern realities of running a school (Levine, 2005; AACTE, 2001; Elmore, 2000). Peterson (2002) argues that professional development for principals holds promise for addressing some gaps in principals' preservice preparation. There are increasing numbers of available professional development programs for principals and an emerging consensus about the ideal qualities and content of such programs (Davis, Darling-Hammond, LaPointe, and Meyerson, 2005). However, there is very little research on the effectiveness of professional development programs for principals (LaPointe Meyerson, and Darling-Hammond , 2006).

At the same time critics have debated the quality of principal preparation, a debate has ensued among educational researchers and policymakers about the role of "scientifically-based" research in evaluating the efficacy of educational practices and programs. As part of this debate, some researchers and federal policymakers have advocated for a greater use of randomized experiments. For example, through the NCLB legislation and the legislation authorizing the Institute of Education Sciences (IES), the federal government has encouraged greater use of randomized experiments to determine the efficacy of educational practices and programs (NRC, 2002; Eisenhart and Towne, 2003). Among the educational research community, randomized experiments have both strong advocates (see for example Slavin, 2002; Boruch, 2002; Cook, 2002) and strong critics (see for example Howe, 2005). Despite the robust debate, which in large part turns on whether experiments should hold a special place of prominence among a broader range of possible research approaches, there remains considerable support among researchers and policymakers that this research approach can provide strong causal evidence of program effectiveness.

This paper reports on the implementation of an experimental evaluation of a comprehensive professional development program for principals called the National Institute for School Leaders (NISL), a program of the National Center on Education and the Economy (NCEE). We begin by briefly discussing the gap in experimental evidence on the principalship. We then discuss four common difficulties in implementing experiments that have been documented in the literature on randomized trials. Next, we describe the design of the NISL experiment, discussing how this design attempted to anticipate these common difficulties. Finally, we discuss the implementation of the NISL experiment, describing problems that were encountered and strategies that were adopted to address the problems. As we report below, randomized experiments focusing on the principalship are virtually nonexistent. We believe that critically examining the use of randomized trials with a population that has heretofore not been the focus of such trials will be informative in its own right, and may also shed light on general issues affecting experiments conducted in educational settings.

Experiments Involving Principals

Against the backdrop of increased advocacy for using experiments in education research, and a widespread belief in the need to improve principal preparation, we sought to assess the available experimental evidence on principals or programs that target principals. 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

(ERIC), Sociological Abstracts, and Criminal Justice Abstracts) and 48 social science journals (Petrosino, Boruch, Rounding, McDonald, and Chalmers, 2000).

Our search of C2-SPECTR identified only 3 manuscripts that focus on principals in some manner. We searched C2-SPECTR using the terms “principal” and “leadership.” These searches identified a total of 18 research articles. Of these, only three articles 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 a 5 day “training laboratory.” The study found “more positive change by principals in the experimental group than by those in the control group, and showed that laboratory training in interpersonal relations affects positively the administrator’s behavior with his staff” (Thomas, 1970). In the remaining randomized experiment, principals were randomly assigned to one of four conditions: 1) principals but not their teachers participated in a school improvement workshop, 2) both teachers and principals participated, 3) teachers participated but principals did not, and 4) neither teachers nor principals participated in the workshop (Grimmet and Crehan, 1987). The results of this experiment indicated that supervision behavior was more effective when both teachers and principals participated in the workshop training.

We note that, similar to C2-SPECTR, the What Works Clearinghouse and other data bases reviewing and abstracting educational interventions such as the Center for Data-Driven Reform in Education (www.cddre.org/), do not include any studies of principals or the effects of

professional development programs for principals. We conclude from this search that experimental evidence on principals, their practice, and the effect of principal training programs (either professional development or preservice) is virtually nonexistent.

Common Problems in Implementing Experiments

There is a rich literature documenting common problems faced by researchers when implementing randomized experiments. In this section we discuss four such problems that were experienced when conducting the NISL experiment – the subversion of random assignment, failure to deliver the treatment as planned, changes in the policy context, and non-compliance with treatment protocols.

Many different aspects of an experimental design can be adversely affected during the implementation of the design, but perhaps no aspect is more important than the randomization of subjects. Boruch (1997) describes how randomization can be subverted at multiple points during the course of a study. The randomization process can be subverted at the outset of the study at the point where subjects are randomly assigned to treatment conditions. Boruch (1997) reports on a medical study in which random assignment was left to hospital admissions staff. These staff members subverted random assignment by assigning patients who should have been in the control group to the treatment group. The hospital staff members switched group assignments because they believed that patients assigned to the control group would benefit from the experimental treatment. Boruch (1997) argues that the best way to avoid this kind of subversion is for researchers to exercise as much control over the process of random assignment as is feasible.

A second common problem in implementing social experiments is that the treatment is not delivered to subjects in the intended manner (Boruch, 1997; Rossi, Lipsey and Freeman,

2004). The treatment that is delivered may differ from the intended treatment in both quantity (i.e. more or less treatment than planned is delivered), and quality (i.e. the treatment is delivered in a different manner than called for by the treatment protocol). Treatment delivery can be hampered by technical, practical and logistical issues. For example, Hirschel et al (1991) found that randomly-assigned police procedures were sometimes not sufficiently strong to ameliorate domestic violence situations, and consequently procedures other than those that were randomly assigned had to be used. Boruch (1997) describes an experiment in which nurses provided an experimental oxygen treatment to infants who had been assigned to the control group, believing that the treatment would be beneficial to these children. Later studies revealed the treatment to be detrimental in some cases.

A third very real problem in experiments conducted in real-world settings is that treatment delivery is also 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 (Rossi, Lipsey, and Freeman, 2004, page 22).

In the case of educational programs such as NISL, the degree to which actors in schools embrace and implement such programs will greatly depend on the extent to which the school district endorses and supports the program. Rossi, Lipsey, and Freeman (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 the 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.

A fourth common problem in the implementation of experiments occurs when research participants fail to follow treatment protocols. Bloom (2005) provides a classification scheme describing different forms of non-compliance with treatment protocols. “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 the treatment. Failure to receive a treatment can occur for a variety of reasons such as subject refusal or because the treatment was not made available to the subject. “Crossovers” are research subjects who are assigned to the control condition but who receive treatment. Again, such subjects can receive treatment for a variety of reasons—because they sought it out, because those responsible for administering the treatment felt they needed it as in the examples above, etc.

The degree to which subjects opt out of treatment can vary from their receiving no treatment to receiving partial treatment. The potential effect of many kinds of treatments, medical, social, and otherwise, clearly depend upon treatment dosage. Consequently, non-participation and non-compliance can greatly diminish a researcher’s ability to detect treatment effects. Boruch (1997) argues that it is important to document treatment delivery as carefully as possible, thereby measuring the amount or dosage of treatment received. Bloom (2005) describes a number of analytic strategies for estimating treatment effects when subjects fail to receive treatment and when subjects who were not intended to receive treatment receive it anyway. Some of these strategies are discussed below.

The NISL Treatment and the Experimental Design

The NISL program is designed to teach principals the theory and practice of instructional leadership within a standards-based policy context. The main goals of the program are to develop principals' abilities to: (1) understand what is entailed in providing high-quality math and literacy instruction; (2) understand what kinds of supports, incentives, and learning opportunities teachers need to improve their math and literacy instruction; and (3) understand what kinds of knowledge and practices they need to employ to lead efforts to improve math and literacy instruction in their schools. The NISL program utilizes face-to-face instruction in workshops, seminars, study groups, and an interactive web learning tool. Educational experts are prominently featured in the curriculum.

The NISL program follows a “train the trainer” model whereby local district “leadership teams” first learn the NISL curriculum and then teach it to local principals. Leadership teams typically include a project director, principals at each level (elementary, middle, high school), district administrators in curriculum and instruction, and possibly local university faculty members. The teams are taught by NISL staff and typically go through two summer institutes and four three-day institutes. In addition to conducting the institutes for leadership teams, NISL coaches also provide substantial post-institute technical assistance. During a district's second summer in the program, local leadership teams begin to deliver the NISL curriculum to principals. This locally-provided training continues for two years with two three-week summer institutes and four three-day institutes for local principals.

The NISL experiment described here is part of mixed-method longitudinal study that was conducted in Cloverville (a pseudonym), a mid-sized urban school district in the Southeastern United States. The original research plan was to use a delayed-treatment experimental design in

which one group of principals was randomly assigned to participate in NISL at the outset of the study (early-treatment group), and a second group was randomly chosen to begin NISL one year after the first group (delayed-treatment group). As discussed below, our design became a simple randomized trial 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 leadership team, all remaining 48 principals in the district were randomly assigned to either the early-treatment or late-treatment groups. We used a basic random assignment design that incorporated school level (i.e., elementary, middle, high) as a blocking variable. In order to prevent the kind of subversion of the randomization process discussed above, a member of the research team performed the random assignment. After random assignments were made, the randomization process was checked by comparing early and late-treatment principals on a wide range of variables measuring school and principal characteristics including gender, race, years of experience and whether the school had met Academic Yearly Progress. These comparisons demonstrated that principals assigned to the two groups were nearly identical on every variable examined.

As part of the data collection process, we instituted procedures for monitoring treatment delivery both in terms of dosage and fidelity. We attended every NISL institute in Cloverville that took place after principals had been assigned to experimental groups. Our primary purpose in attending the institutes was to collect qualitative data on the nature of the treatment received by Cloverville principals in NISL trainings. We also collected attendance data for each training that we later used to document those principals who were assigned to receive the NISL treatment but failed to do so ("no shows") and those principals who were not supposed to attend NISL

trainings but did so ("crossovers"). This also allowed us to monitor whether participants attended the sessions in full or only attended parts of the training during a given session.

We also instituted procedures for retaining sample members over the three years of the study. Boruch (1997) argues for the importance for knowing subjects' whereabouts over the course of a longitudinal experimental study. During each year of the experiment we had multiple points of contacts with Cloverville principals. Careful records were maintained when principals either left the district or changed schools within the district. Boruch (1997) also argues for the importance of providing financial incentives to participants when participation is voluntary in order to retain sample members. The NISL study had multiple components and principals received a separate incentive payment for participating in each component. Principals who participated in all study components received incentive payments totaling \$235 per year.

A Chronology of NISL Implementation and the Experimental Study

In conducting the experimental evaluation of NISL we experienced each of the four challenges previously discussed—the subversion of random assignment, problems with treatment delivery, shifts in policy, and non-compliance of research subjects. In this section we present a chronology of how the NISL training and the evaluation study unfolded. We first discuss how changing district priorities impacted the delivery of the NISL program. We then describe levels of non-compliance with the NISL program and discuss how random assignment was subverted.

Context Factors that Affected Treatment Delivery

Before it began, the NISL study was affected by the kinds of policy shifts identified by Rossi, Lipsey, and Freeman (2004). Initially, our proposed research was to be conducted in Brockville (a pseudonym), which is among the 20 largest school districts in the U.S. At the time we proposed the NISL study, the Brockville district had already developed a leadership team and

had begun piloting the NISL curriculum. But despite having a long relationship with NCEE, Brockville chose not to expand NISL beyond the initial group, opting instead to use their own leadership development program.

Upon learning that Brockville was no longer offering NISL, we were able to successfully negotiate with the Cloverville district to serve as a site for the NISL randomized trial. 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 3-4 districts across the country that had adopted the NISL program and was therefore the most suitable research site at the time. In addition, we had to locate a district that had not yet begun the NISL program so we could implement random assignment and pre-testing.

During the summer and fall of 2004 the Cloverville leadership team participated in institutes provided by NISL staff. Most of the leadership team was enthusiastic about NISL, and the 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 NISL experiment was selected and principals were randomly assigned by members of our research team to early-treatment and late-treatment groups as described above. During the spring of 2005 we collected baseline measures on principals and teachers. In the summer of 2005, members of the leadership team participated in another institute provided by NISL staff.

In 2005, the study continued to be affected by shifts in district priorities. In the spring and summer of 2005, Cloverville hired a new superintendent who began in the Fall of 2005. The prior superintendent, Mr. Anderson (pseudonym) was largely responsible for bringing NISL to the district. He had a fairly longstanding relationship with NISL staff prior to coming to Cloverville, and saw NISL as the primary vehicle for developing school leaders in the district. In

addition, the director of professional development was a key liaison to our research team. She was also very knowledgeable about NISL and was responsible for its ongoing scheduling and implementation in the district.

The new superintendent, Mr. Johnson (pseudonym), had no prior experience with the NISL program and brought his own ideas and preferences for leadership development with him into the district. The director of professional development left soon after the new superintendent was hired to pursue an opportunity in another larger urban district. One major consequence of the arrival of a new superintendent was that district leadership decided that NISL training would *not* be provided to the second cohort of principals (i.e. the late-treatment group). From the perspective of the study, that meant that the design was no longer a delayed-treatment design, but instead, was a simple randomized trial, since the early-treatment group was now the only group of principals who would receive treatment. In addition, we have learned that the new superintendent had his own professional development initiatives for school principals. He began to bring to Cloverville some of the training that he had used in his previous district and he showed little engagement with NISL.

Principals were originally intended to receive two years of NISL 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 2 year course of NISL training, but that principals in the late-treatment group would not begin the program. Had this plan been enacted, early-treatment group principals would have continued to receive NISL training during the 2006-2007 school year. As of this writing (March, 2007), no NISL training had been provided during the 2006-2007 school year. Interviews with district leaders conducted in February of 2007 provided contradictory evidence about whether additional NISL training would be provided to the early-

treatment group. In light of the history of the NISL program in the district, however, we are doubtful that any additional training will be provided.

Furthermore, given the changes in the district context, it is also likely that the fidelity of the treatment was compromised. Treatment fidelity could be vulnerable to changes in district context with a program like NISL because it is a train the trainer model. Local leaders in the district were trained to implement the NISL program and curriculum. However, when the superintendent changed, these local leaders have to attend to the new professional development priorities of the new superintendent simultaneous to their implementation of NISL.

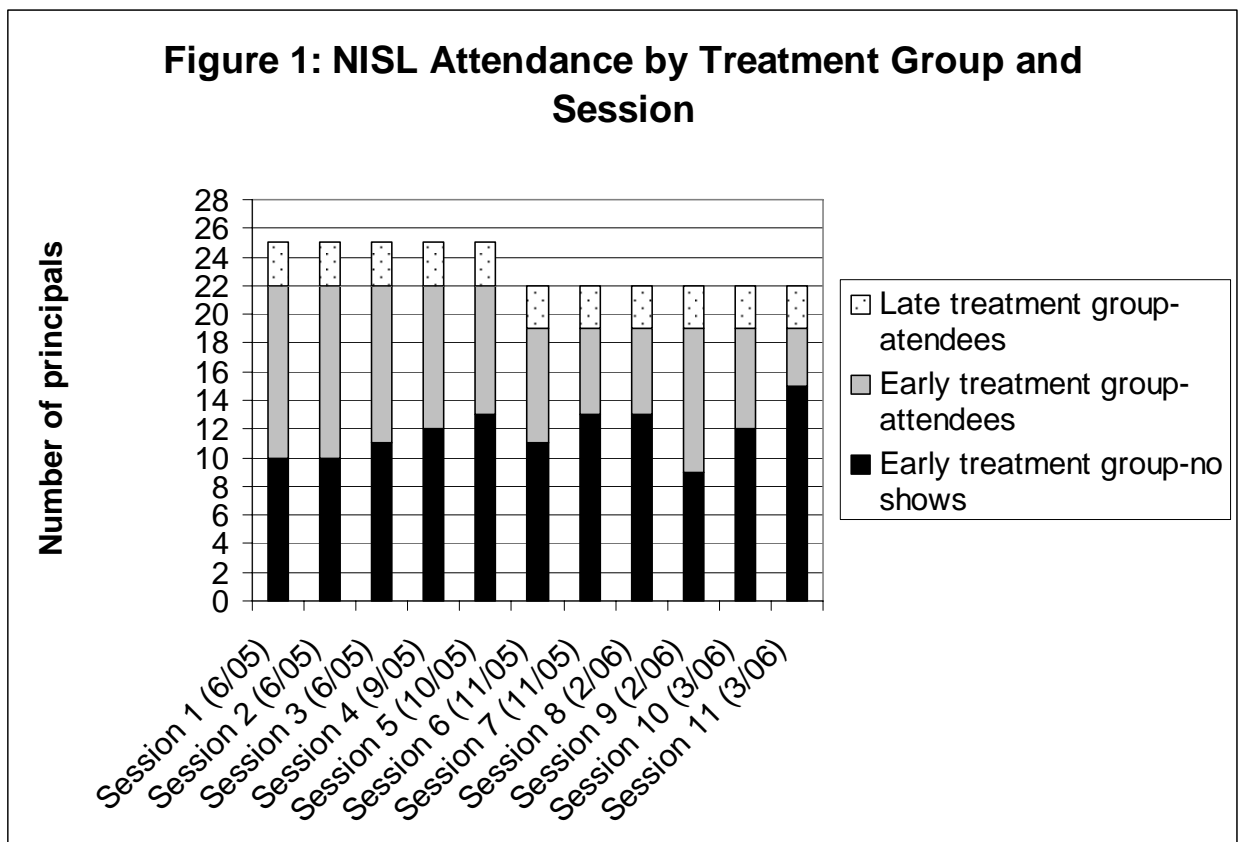
Subversion of Randomization and Non-Compliance

In June of 2005, members of the leadership team delivered the first units of the NISL curriculum to principals. At this initial training, we found evidence of the subversion of treatment assignment and non-participation. Of the 24 principals assigned to the early-treatment group, only 12 attended the first NISL training in June. This pattern continued as 10 of these 12 principals did not attend a single NISL training. While only half of the principals who were expected to attend training did so, three principals who were *not* assigned to attend NISL training did so. Based on the qualitative data we have collected to date, it is not clear whether the subversion of random assignment occurred at the district level or whether these three principals simply sought out the NISL training on their own accord. We have plans to conduct follow up interviews with these three “crossover” principals in order to learn more about this issue.

During the 2005-2006 school year, an additional 5 units of the NISL curriculum was delivered by members of the leadership team. Similar patterns of attendance evident at the first session in June were observed in subsequent training sessions, with three principals assigned to the late-treatment group regularly showing up, and substantial numbers of principals assigned to

the early-treatment group failing to show up. A total of eleven NISL training sessions were delivered to Cloverville principals. Figure 1 displays the attendance for each session.

We note that overall attendance at these sessions decreased over time, as did the number of eligible principals in the treatment group. Principals assigned to the early-treatment group had an overall attendance rate of about 42%. Clearly the poor attendance at NISL training sessions will lessen our chances of observing treatment effects for the program. Our chances are even further diminished because the district has only delivered half of the planned NISL training to the early-treatment group.



This chronology of events shows how we experienced a number of common problems that are documented in the literature on the implementation of experiments (Boruch, 1997; Bloom, 2005; Rossi, Lipsey, and Freeman, 2004). The study was adversely affected by major shifts in district priorities. When we began the study, Cloverville was strongly committed to the NISL program, and, as we learned through interviews from that time, there was a great deal of optimism in the district that the program would bring about positive change in principal leadership. During the early months of program implementation, the NISL program was both supported and carefully managed by senior district leadership. However, after the new superintendent arrived and the former director of professional development departed, the program no longer enjoyed the support and attention of senior management and was subsequently allowed to essentially ‘whither on the vine.’ The end result was that only about half of the treatment was delivered to the early-treatment group and none of the treatment was delivered to the late-treatment group.

We also experienced problems commonly associated with research subjects’ participation in experiments. The random assignment process was subverted as three principals who were assigned to the late-treatment group attended all eleven of the NISL training sessions provided in Cloverville. The fact that there were three “crossover” principals, and that there was great consistency in the attendance of these three principals suggests to us that the district may have been responsible for the subversion of randomization by encouraging these principals to attend trainings. To date, however, we do not have solid evidence about whether this occurred, or whether these three principals were acting independently. We also observed weak attendance at NISL training sessions. The average attendance rate for early-treatment principals who were

eligible to attend NISL trainings was 42 percent. In addition, a total of ten principals assigned to this group did not attend a single NISL training.

Using Local Average Treatment Effects (LATE) to Assess the Impact of Treatment on the Treated

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 thought to be affected by program participation.¹ This comparison, which is based on subjects' assignment to treatment, reflects the *intended* treatment plan, often referred to as the *intent to treat (ITT)*. Boruch (1997) argues that there is both a scientific rationale and a policy rationale, 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 for conducting ITT analyses rests on the fact that groups that have been formed via random assignment do not systematically differ from one another. Consequently, comparisons of randomly-created groups will produce unbiased estimates of treatment impact which allow for legitimate causal statements about treatment effects. Boruch (1997) further argues that ITT analyses also have relevance to social policy, 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 addresses the policy question: “What is the likely effect of this treatment if it was *made available* to people like those studied?”

People are often interested in a different question about program effectiveness—Did the program have an impact on people who participated in it? In the literature on experimental

¹ In the current experiment, the primary outcomes of interest are pre-treatment and post-treatment measures of principal practice collected with annual surveys and daily logs.

research, analyses that address this question are said to assess the effect of *treatment on the treated* (TOT). Drawing upon the work of Angrist, Imbens, and Rubin (1996), Bloom (2005) outlines a strategy for estimating a type of TOT effect when persons assigned to the treatment group “crossover” and receive the treatment, just as three principals did in the NISL experiment. Bloom (2005) and Angrist et al (1996), refer to this estimator as a local average treatment effect (LATE). LATE is calculated by adjusting the ITT effect (i.e. the simple mean difference between treatment and control groups) for rates of treatment participation among treatment and control groups. The formula for calculating LATE is as follows:

$$LATE = \frac{Y_T - Y_C}{P_T - P_C}$$

where...

Y_T is the mean for subjects assigned to the treatment group

Y_C is the mean for subjects assigned to the control group

P_T is the proportion of subjects assigned to the treatment group who received treatment

P_C is the proportion of subjects assigned to the control group who received the treatment

The formula for LATE essentially “inflates” the size of ITT effects based on the total number of people who received treatment, and therefore, the degree to which LATE estimates exceed ITT estimates increases as the overall number of people who experienced treatment increases. According to Bloom (2005), LATE represents the average effect of treatment for a person who was induced by randomization to receive the treatment.

When analyzing the impact of the NISL program among Cloverville principals we will examine both ITT and LATE estimates of treatment effects. Examining the ITT effects on outcomes that are hypothesized to be sensitive to the NISL program will provide a realistic “policy” perspective about the kinds of results districts might expect under realistic conditions.

Of course, the conditions in Cloverville might be fairly unique, so we are equally interested in trying to assess the impact of the program under more ideal conditions. Examining LATE effects will allow us to get a sense of how much stronger the effect of NISL might be if a greater number of principals had participated.

Conclusions

In conducting an experimental evaluation of an executive professional development program for principals, we experienced many of the challenges documented in the literature on the use of experiments in the social sciences. In the course of our experiment, policy changes resulted in the program not being adopted in the original study site. In the second district, leadership changes transformed the program in its early stages from a favored reform to a peripheral one. In the wake of these shifting district priorities, substantially less treatment than planned was delivered to principals, those assigned to receive the treatment participated at low rates, and a small number of principals who were not assigned to receive the treatment participated in the intervention. The former two conditions resulted in a greatly reduced “dose” of the NISL treatment being delivered to Cloverville principals, and the reduction in treatment dose substantially reduces the likelihood that we will observe an effect of the NISL program in Cloverville. In our view, the complex treatment conditions in Cloverville might not be understood very well through a basic intent to treat analysis. Consequently, we are currently exploring alternative ways of understanding treatment effects, including the use of quasi-experimental statistical analyses, and the analysis of qualitative data collected from NISL participants and district staff.

Our experiences with conducting the randomized trial in Cloverville reinforce the importance of documenting treatment delivery and treatment receipt. Research staff members

attended every NISL training session provided to principals beyond the initial institutes for members of the leadership team. For every principal assigned to the two experimental groups we have a record of whether or not they attended each of the eleven sessions that were delivered. From this database we are able to determine whether a principal attended any trainings, how many trainings a principal attended (a potential measure of treatment dosage level), and whether or not a principal's attendance was a subversion of random assignment. Such variables will allow us to model treatment effects much more accurately.

Our experience in conducting the NISL experiment also reinforced the usefulness of using mixed method designs to go beyond the main effects of the treatment and to probe how those effects come about. We have collected qualitative data on treatment delivery and participation through interviews with NISL participants and trainers. We have also collected qualitative data that will allow us to probe more deeply into program impact. In future research, we plan to use qualitative data to gain an understanding of the *quality* of treatment delivery, how participants experienced the treatment, what they learned, and how they changed their leadership practices.

In conducting the experimental evaluation of NISL we were also confronted with questions about the practicality and validity of using randomized trials to evaluate large-scale educational programs in complex, regularly-changing settings. Newmann, Smith, Allensworth, & 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 Fred 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 NISL 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 (Rossi, Lipsey, and Freeman, 2004; Borman, Slavin and Cheung, 2005). However, our experience in conducting the NISL experiment suggests that scouting out amenable contexts is not foolproof. As discussed above, both Brockville and Cloverville were both very supportive of NISL when we initially approached these districts. However, in both settings, support for NISL eroded.

Finally, though we have limited direct evidence, we conjecture that principals may be a particularly challenging population to get to comply with treatment assignments. As the chief executive officers of their schools, which are sometimes very large, complex organizations, principals are busy, autonomous individuals. Programs like NISL, and the studies that examine their effects, depend on the voluntary participation of these busy, autonomous executives. As of this writing, we have not been able to completely sort out the extent to which low attendance rates among principals assigned to the treatment group were driven by constraints on principals' time, principals' lack of interest in the NISL curriculum, encouragement or discouragement from district staff. Clearly there is a dearth of experimental evidence on principals and professional development for principals. However, future experiments targeting this population would be well advised to explore ways to encourage participation in programs to which principals are randomly assigned, and to seek out research sites which 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.

References

- American Association of Colleges for Teacher Education. (2001, March). PK-12 educational leadership and administration (white paper). Washington, DC.
- Angrist, J., Imbens, G., and Rubin, D. (1996). Identification of causal effects using instrumental variables.” JASA Applications invited paper, with comments and authors’ response. Journal of the American Statistical Association. 91(434), pp 313-335.
- Borman, G.D., Slavin, R.E., and Cheung, A. (2005). Success for All: First-Year Results from the National Randomized Field Trial. Educational Evaluation and Policy Analysis, v27 n1 p1-22
- Boruch, R. (2002). The Virtues of Randomness. Education Next. 2(3), pp. 36–42
- Boruch, R. (1997). Randomized experiments for planning and evaluation: A practical guide. Thousand Oaks, CA: Sage.
- 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(3), pp. 175–199.
- Davis, S., Darling-Hammond, L., LaPointe, M.A., and Meyerson, D. (2005). *School Leadership Study: Preparing Successful Principals. Review of Research*.
- Eisenhart, M and Towne, L. (2003). Contestation and change in national policy on “scientifically based” education research. Educational Researcher. 32(7), pp 31-38.
- Elmore, R. F. (2000, Winter). *Building a New Structure for School Leadership*. Washington, DC: The Albert Shanker Institute.
- Heath, L., Kendzierski, D., & Borgida, E. (1982). Evaluation of social programs: A multimethodological approach combining a delayed treatment true experiment and multiple time series. Evaluation Review, 6(2), 233-246.
- Hess, F.M. (1998). *Spinning Wheels: The Politics of Urban School Reform*. Washington DC: Brookings Institution Press.
- Hirschel, J.D., Hutchison, I.W., Dean, C.W., Kelley, J.J., & Pesackis, C.E. (1991). Charlotte Spouse Assault Replication Project: Final report. Charlotte: University of North Carolina at Charlotte.
- Howe, K.R. (2005). The question of education science: experimentism versus experimentalism. Educational Theory. 55(3), pp 307-321.

- LaPointe, M., Meyerson, D. & Darling-Hammond, L. (2006). Preparing and Supporting Principals for Effective Leadership: Early Findings from Stanford's School Leadership Study Paper was presented at the 2006 annual meeting of the American Educational Research, San Francisco, CA
- Levine, A. (2005). *Educating School Leaders*. New York: The Education School Project.
- 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*, v23 n4 p297-321
- National Research Council (2002). *Scientific research in education*. R. Shavelson and L. Towne (Eds.), Committee on Scientific Principles for Educational Research. Washington DC: National Academy Press.
- Peterson, K. D. (2002). The professional development of principals: Innovations and opportunities. *Educational Administration Quarterly*. 38(2), 213-232.
- Petrosino, A., Boruch, R.F., Rounding, C., McDonald, S., and 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(3&4), p206-19.
- Rossi, P.H., Lipsey, M.W., Freeman, H.E. (2004). *Evaluation: A Systematic Approach*. Thousand Oaks, CA: Sage.
- Thomas, T.A. (1970). *Changes in elementary school principals as a result of laboratory training*. University of Oregon, Center for Advanced Study of Educational Administration.
- Young, I.P. and others. (1997). Holmes versus Traditional Teacher Candidates: Labor Market receptivity. *Journal of School Leadership*, 7(4), p330-44.