Contract No.: ED-04-CO-0112/0002
MPR Reference No.: 6212-200

MATHEMATICA
Policy Research, Inc.

Options for Studying Teacher Pay Reform Using Natural Experiments

March 30, 2006

Mathematica Policy Research, Inc.
Steven Glazerman
Tim Silva
Nii Addy
Sarah Avellar
Jeffrey Max
Allison McKie
Brenda Natzke

Chesapeake Research Associates
Michael Puma
Patrick Wolf
Rachel Ungerer Greszler

Submitted to:

U.S. Department of Education, IES/NCEE
555 New Jersey Avenue, N.W., Room 500H
Washington, DC 20208
Telephone: (202) 219-2129
Facsimile: (202) 219-1725

Project Officer:
Stefanie Schmidt

Submitted by:

Mathematica Policy Research, Inc.
600 Maryland Ave. S.W., Suite550
Washington, DC 20024-2512
Telephone: (202) 484-9220
Facsimile: (202) 863-1763

Project Director:
Steven Glazerman

ACKNOWLEDGMENTS

Mathematica Policy Research, Inc., and Chesapeake Research Associates are grateful to the many individuals who generously provided guidance and other assistance on both the feasibility analysis and this report. Stefanie Schmidt, our project officer at the U.S. Department of Education (ED), provided input at each step, as did Betsy Warner, also of ED. Numerous officials in district and state education agencies as well as those in data centers in Florida, North Carolina, and at the Northwest Evaluation Association and the Teacher Advancement Program Foundation were very helpful and generous with their time over the phone and by email. We also thank Lynn Cornet of the Southern Regional Education Board, Mike Podgursky of the University of Missouri-Columbia, and Garry Ritter and Jay Greene of the University of Arkansas.

At Mathematica, we are grateful to John Deke and Paul Decker, who carefully read drafts of the report, as well as Molly Cameron, who edited the report, and Donna Dorsey who expertly prepared the manuscript.

CONTENTS

A. BACKGROUND: STUDYING TEACHER INCENTIVES

How public school teachers are paid in the U.S. has been a controversial issue for many years. Critics of the traditional system, in which teachers are paid solely on the basis of years of experience and educational attainment, claim that it does not reward or promote good teaching as fairly as systems, for example, that tie pay to performance, having certain skills, or being willing to teach in high-need areas. Proponents of the traditional system argue that experience and education are important predictors of teacher performance, and that the simplicity, transparency, and fairness of the system make it the only logical choice. In an attempt to achieve the best of both worlds, educators and policymakers have devised numerous approaches to reforming the teacher payment system, each of which seeks to fine-tune teacher incentives in different ways.

Choosing between the various approaches is difficult, however, because the scientific evidence on their effectiveness is extremely limited. History has shown that successfully *implementing* a teacher pay reform policy, much less conducting a rigorous study of one, is a formidable challenge. Many of the most ambitious and interesting reforms have collapsed within a few years under pressure from political opposition or fiscal constraints, and attempts to study the few reforms that stayed afloat have yielded little fruit to date (Glazerman 2004).

Recognizing these challenges, the U.S. Department of Education (ED), Institute for Education Sciences (IES) has contracted with Mathematica Policy Research, Inc. (MPR) and its subcontractor, Chesapeake Research Associates (CRA), to explore the feasibility of analyzing existing education data—typically from states and school districts—to provide quasi-experimental evidence of the promise of alternative teacher compensation strategies or incentive plans. The quasi-experimental evidence comes from natural experiments, where eligibility for a teacher incentive program varies for some reason that is not related to outcomes. Examples include variation over time, where the timing of the program is arbitrary, variation across districts, where eligibility rules arbitrarily disqualify districts whose enrollment is just over some arbitrary threshold, or variation across schools and time, where the timing of program adoption was more or less random. The primary research question that the secondary data analysis will address is the following: Do teacher incentive programs improve student learning, either by making teachers more productive (a productivity effect) or by attracting and retaining good teachers at higher rates (a composition effect)?

This report presents findings and recommendations from our review of how secondary data could be used to study a variety of teacher incentive programs. We conducted an extensive search for alternative pay programs, identifying more than 100 possibilities and, in consultation with IES, narrowed the list to seven programs that are the most promising for evaluation. The next section (Section B) discusses the preliminary steps of identifying and categorizing teacher incentive programs and the process of narrowing down the list to the most promising. Section C provides detailed profiles of seven candidates for further study. Sections D and E discuss some cross-cutting issues of sample size adequacy and cost, respectively. Finally, Section F summarizes the findings and offers recommendations for next steps.

B. TYPOLOGY AND LANDSCAPE OF TEACHER PAY PROGRAMS

1. Categorizing Teacher Incentive Programs

For this feasibility study, we grouped teacher pay reform strategies into three categories according to what is being rewarded. Alternative teacher pay plans vary along many other dimensions as well, including the method for measuring teacher performance and the size of the bonuses (Glazerman 2004). However, we found that the following three categories would be the most useful way to list candidate programs for the proposed study for IES:

1. ***Pay for Performance***. These pay plans reward teachers for performance, which is measured by student achievement or other data, such as a supervisor's (principal's) rating or a score from a formal classroom observation protocol.
2. ***Pay for Knowledge, Skills, or Extra Responsibility***. These pay plans reward teachers who demonstrate special skills or knowledge, or who take on extra responsibilities associated with improvements in classroom instruction or student achievement. These pay plans, unlike performance-based pay, do not link compensation to measurements of student achievement or to teacher performance in the classroom.
3. ***Pay for Filling a Need***. These pay plans use incentives to attract or retain qualified teachers for shortage areas, which may be defined by geography (e.g., rural areas), student characteristics (e.g., high poverty) or subject area (e.g., high school chemistry).

There is considerable variation within each category and the groups are not necessarily mutually exclusive. Many programs provide multiple ways for teachers to earn bonuses, but we found this categorization to be an effective way of distinguishing study opportunities that would call for different design or measurement strategies and emphasize different outcome variables.

2. Mapping the Landscape of Programs and Study Opportunities

As mentioned above, we conducted an extensive search and identified more than 100 potential candidates for in-depth study.[1] We began with several lists of known teacher incentive and pay reform programs such as those mentioned in Glazerman (2004), Cornet and Gaines (1991, 2002), Solmon et al. (2005), and Hassell (2002) and supplemented this with electronic searches in the *Education Week* archives, state departments of education websites, and other academic and popular online search engines, as well as extensive communications with

[1] Not all of the pay plans in our database were stand-alone programs, but in this report we will call them "programs." In some cases we combined a number of pay plans, which we then listed as a program (e.g., the Teacher Advance program [TAP]). In other cases, where a pay plan was being implemented by one organization, we listed a plan as a separate program if it had distinctly different components for different target groups (e.g., Charlotte-Mecklenburg, NC).

policymakers and other researchers who study teacher compensation. We conducted searches that focused on each of the 50 states and the District of Columbia and populated a database with teacher pay programs at the federal, state, and local levels. We used the database to track several pieces of information relevant to the task of identifying promising candidates for further study, including: what is being rewarded, the types and sizes of rewards, the form of rewards, the mechanisms used to make award decisions, who qualifies for rewards, and the visibility of the program. We also noted the time periods (years) during which a program was implemented, the data one could obtain to study it, and the availability of any previous studies of the program.[2]

3. Narrowing the List of Candidates

We selected 20 potential study candidates from our database by eliminating a large number of programs that did not meet our feasibility criteria, which included policy relevance, availability of data, and opportunity for a plausible quasi-experimental design. Where necessary, we contacted program officials for additional information, focusing on the following details about programs in our database:

- ***Policy Relevance.*** For each of the identified programs, we tried to verify whether it was implemented as intended. We also noted how widely each program has been implemented and how feasible broader implementation (replication) may be.
- ***Data Availability and Accessibility***. We noted which outcomes could be studied with the data.

Our additional searches and phone calls to state and local education agencies and school districts revealed that a large number of programs that were listed on websites or other media were not feasible study candidates for reasons that included the following:

- ***Weak or Late Implementation***. Many programs were not as widely implemented as we initially had believed or were implemented on a slower schedule or in different ways from the plan outlined in legislation or district website descriptions. Many programs, such as the well-known Denver Pay for Performance program, were implemented too recently to be included in a study that can be completed in a timely manner. A number of programs were widely publicized, but had not been implemented at all for various reasons, including lack of funding or political opposition.

[2] The database on teacher pay programs was not filled in completely. Once we identified characteristics that made a program infeasible to study—for example, if the program had not been implemented until the 2005-06 school year—we did not continue to gather additional information on other factors, such as data sources or detailed program rules.

- ***Lack of Data***. Some states had not collected the data that would be needed to study the programs. In other cases, test scores and other relevant data were available for only part of the period that one would need to study a program.

- ***No Comparison Group or Natural Experiment***. Some programs had been universally implemented, and it would have been difficult or impossible to identify a valid comparison group. For such programs, it would have been challenging to identify program effects.

In consultation with IES, we selected the seven most promising candidates from our list of 20 so as to glean important insights into the effectiveness of teacher pay reform. Where necessary, we verified some of the information about the 20 programs (notably about participation in the programs and availability of data). We then chose these seven programs to represent each category of teacher pay reform strategies with at least one example. Following IES's recommendations, we selected three programs rewarding performance; two programs rewarding knowledge, skills, or extra responsibilities; and two focusing on recruitment and retention (addressing a need). The details of these programs are presented in Section C.

4. Candidates Considered but Not Recommended

Here we briefly describe two programs that *appeared* promising and the reasons we excluded them from the list of promising programs after conducting additional background research. In an earlier draft of this report we identified and profiled eight promising programs. However, in updating the preliminary information for all eight, we decided to remove two programs from the list—the Benwood Initiative in Hamilton County (Chattanooga), Tennessee, and the Douglas County Pay for Performance Program in Colorado—and add one new program—the Palm Beach County Title I Sign-On Bonus Program.

The **Douglas County Performance Pay Plan** was on our initial short list as promising, but we removed it because the research design was not as strong as the other options we identified. The Douglas County plan has received national attention because of its longevity and because it differs from most salary schedules by rewarding annual pay increases for years of satisfactory experience. The county has continuously implemented the pay plan since 1994, with minor adjustments and additions over time. A distinctive feature of the plan is that teachers must receive a "satisfactory" rating on their performance evaluation to earn an annual length-of-service salary increase. Teachers also have the opportunity to earn six different bonus incentives each year, including a one-time payment of $1,250 for the Outstanding Teacher Bonus, and $12,500 over five years for the Master Teacher bonus. The Outstanding Teacher bonus is based on a teacher portfolio submission. To earn the Master Teacher designation, a teacher must demonstrate student growth, professional leadership, and professional recognition. A paper by the Mid-continent Research for Education and Learning (McREL) found that the Douglas County pay plan did not affect recruitment of quality teachers, but did reduce attrition among high quality teachers (Reichardt and Van Buhler 2003).

Because the Douglas County pay plan is a district-wide program, we would be constrained to a district comparison design that would compare student achievement growth in Douglas County with growth in comparable Colorado school districts that did not implement an

alternative pay plan. Since only one district implemented the program we can only estimate the "Douglas County Effect" and attribute it to the district's unique way of paying teachers. To the extent that any other characteristics of Douglas County are unique relative to other districts in the state, we would be confounding those features with the pay plan. The pay plan has been operating for a long time, so there is limited opportunity to use pre-implementation outcomes as a benchmark because the data are not available for the years before 1994. Another potential downside to studying the pay plan is that it comprises many different components, making it difficult to isolate or identify the effect of any one aspect of the plan. The inter-district comparison, with only one district having implemented the program and no pre-implementation data available, does not provide an adequate basis for measuring the effect of Douglas County's performance pay plan.

The **Benwood Initiative** is a program in Chattanooga, Tennessee that also appeared promising in a number of ways. With a teacher incentive plan as its central feature, the Benwood Initiative was begun in 2002 to improve student achievement, especially reading proficiency, in the district's lowest-performing schools. The program targeted nine schools, all of which are located in the county's urban area, in the city of Chattanooga. The schools were targeted because they had been listed as being among the 20 lowest performing schools in the state. The initiative aims to attract and retain high-quality teachers by offering cash bonuses and other benefits to teachers and other school staff, mostly on the basis of student test-score gains. It is hypothesized that teachers' advance knowledge of the potential bonuses and other benefits will attract them to and/or retain them at Benwood schools, and motivate them to help students make greater annual achievement gains than they might have made otherwise. It also aims to bring about improvements through reconstitution of the schools' leadership and teaching staff and through a large investment of resources for professional development, materials, after school programs, and additional staff.

The Benwood Initiative provides a series of bonuses and performance incentives that can add a large amount to a teacher's compensation. The individual teacher performance incentives include $5,000 bonuses for high scores from the Tennessee Value-Added Assessment System (TVAAS) and eligibility for free enrollment in the Master's degree program in urban education at the University of Tennessee. School teams can earn bonuses of $1,000 or $2,000 for schools whose students achieve three-year average gain scores above a threshold. The team bonuses are given not only to classroom teachers, but also to principals, assistant principals, special subject teachers, and librarians. Principals at schools achieving the team bonuses receive annual lump-sum bonuses of $10,000 in addition to the team bonuses. Beginning with the 2004-05 school year, assistant principals at such schools also receive additional bonuses of $5,000. To increase the retention, performance bonuses for a given school year are paid out in the following year, so awardees must continue to teach in a Benwood school to receive the bonus.

In addition to performance bonuses, all Benwood teachers are eligible for financial help to buy a home in downtown Chattanooga. Teachers can receive a loan of up to $10,000 for a down payment, which will be forgiven if they live in the homes for five years. A second forgivable mortgage of up to $20,000 can be applied to add to the down payment and pay closing costs.

We are not recommending a quasi-experimental evaluation of the Benwood program because it includes many different intervention components beyond teacher bonuses, making it infeasible to isolate the effects of just the offer of teacher incentives. As mentioned above,

Chattanooga implemented a number of measures in its Benwood schools at the same time as the bonus award program. The district began *reconstitution* at approximately the same time that the incentive program started. In 2001-02, the district replaced six of the nine principals at Benwood Schools. Also, teachers were required to reapply to the schools, leading to some voluntary and involuntary transfers of more than 50 teachers out of the schools. In 2001-02, Chattanooga Public Education Foundation (PEF) and the Benwood Foundation of Chattanooga provided grants to provide *professional development* for all Benwood classroom teachers in literacy instruction and instructional strategies for urban learners. Schools in the district also used PEF-Benwood funding to place a wide variety of *books in all classrooms*, hire *reading specialists* to work with struggling readers, provide *coaches for new teachers*, and provide *leadership coaches* to help principals and assistant principals guide and evaluate teachers. The district implemented other improvement strategies, such as *reorganizing the school day* to allow concentrated study of reading and writing, instituting *after-school and summer school programs* for all students, hiring a full-time *parent involvement coordinator*, providing *mentoring programs* for new teachers, and providing special *enrichment activities* for students. This wide variety of interventions would all be perfectly confounded with the introduction of performance incentives for teachers.

C. PROMISING PROGRAMS

After we identified the seven programs for detailed feasibility assessment, we gathered more in-depth information on program details, prior research, and data availability, and we developed an approach to estimating impacts. These pieces of information constitute the organizing principle for the detailed profiles of each of the seven programs below, which are candidates for further study. Table 1 lists the names and locations of each program according to the categories described in Section B.

Each profile follows the same outline. The profiles include the program's goal; any distinctive features that set it apart from other programs we considered; the rules for teachers, schools, or districts to be eligible for awards; the basis for rewarding teachers; the size of rewards; and dates during which the program has operated. We also examine the program in its larger policy and research contexts to determine if evaluating it would have national implications and would add new information beyond the existing literature.

To develop ideas for the study design, we first identify the research question or questions that we hope to answer. We then discuss design options, i.e. the means by which we hope to identify program effects, given that we will be using observational data and not data from controlled experiments. Next, we discuss key variables, means of measurement, and the availability of data needed to conduct the study. Finally, we discuss some advantages and disadvantages of each candidate for further study.

TABLE 1

PROGRAMS IDENTIFIED AS PROMISING CANDIDATES FOR FURTHER STUDY

Program		Location
Pay for Performance		
1.	The Teacher Advancement Program	Multiple states (selected schools in MN, CO, SC, AZ, FL, AR)
2.	Charlotte-Mecklenburg Performance Based Pay Plan	North Carolina (Mecklenburg County)
3.	California's Certificated Staff Performance Incentive Award Program	California
Pay for Knowledge, Skills, or Extra Work		
4.	Cincinnati Teacher Evaluation and Compensation System	Ohio (Cincinnati)
5.	Missouri Career Ladder Program	Missouri
Pay for Filling a Need		
6.	Arkansas High Priority District Bonus Program	Arkansas
7.	Palm Beach County Title I Sign-on Bonus Program	Florida (Palm Beach County)

1. TEACHER ADVANCEMENT PROGRAM (TAP)

a. Overview of the Program

Goal. The Teacher Advancement Program (TAP) aims to attract talented individuals to teaching and retain them in the profession by offering opportunities to earn higher salaries and advance in their careers without leaving the classroom. Under the TAP model, teacher pay and advancement are tied to student achievement growth, observed performance in the classroom, qualifications in high-demand subjects, and a willingness to take on mentoring duties. TAP also seeks to improve teacher quality through ongoing professional development and accountability for performance.

Program Structure and Operation. The Milken Family Foundation based in Santa Monica, California developed TAP as a comprehensive school reform model in the late 1990s. This school-wide program provides teachers opportunities for additional pay for performance (measured through expert observers and analysis of student test score gains), career advancement with corresponding pay raises, and continued professional growth, while at the same time holding teachers accountable for student learning.

The following four principles define the TAP strategy for attracting, motivating, and retaining high-quality teachers:

1. ***Multiple Career Paths*.** Classroom teachers may remain "career" teachers or seek promotion to a mentor or master teacher position. Together with the principal, mentor and master teachers comprise the school leadership team that oversees all TAP activities. Master and mentor teachers receive increased compensation for assuming additional responsibilities, which include supporting the professional growth of other teachers and working with the principal to set achievement goals and evaluate teachers. Promotion to a mentor or master teacher position occurs through a competitive, performance-based selection process. The principal makes the final promotion decision based on input from administrators and a committee of teachers.

2. ***Ongoing Applied Professional Growth*.** TAP builds time into the school week for school-based teacher learning targeted towards addressing identified student needs. Teachers meet in weekly cluster groups led by mentor or master teachers. Each teacher also develops an individual growth plan that includes specific goals and activities. In addition, mentor and master teachers provide other teachers with ongoing classroom support.

3. ***Instructionally Focused Accountability***. Four to six times a year, each teacher undergoes an evaluation by multiple trained and certified evaluators. Using the *TAP Teaching Skills, Knowledge, and Responsibility Standards*, teachers are evaluated both individually, based on the learning growth achieved by the students in a given teacher's classroom, and collectively, based on the learning growth of all students in the school.

4. ***Performance-Based Compensation*.** Teachers may earn annual financial awards based on both instructional performance, as observed in multiple teacher evaluations,

and student achievement growth. Both classroom-level and school-level achievement growth factor into performance pay. TAP also encourages districts to provide competitive pay for teachers in high-need subjects and schools.

The Milken Foundation staff who developed and operate TAP formed the TAP Foundation to provide support for schools in implementing the program. The TAP Foundation offers training and certification services to prepare master and mentor teachers for evaluating other teachers and conducting professional growth activities. In addition, principals may use the TAP Performance Appraisal System in organizing and tracking teacher evaluation data throughout the school year.

TAP emphasizes that performance pay is not a stand-alone component of the program. Rather, differential compensation as reward for increased responsibility, observed instructional performance, and contributions to student achievement is embedded in the defining principles of the program.

Awards. The size of the financial awards available to TAP teachers, which supplement the existing teacher salary scale, varies by school. According to our communications with TAP program staff, the additional compensation for master teachers ranges nationally from $5,000 to $11,000. Bonuses for mentor teachers range from $2,000 to $5,000. The performance bonuses have three components: 50 percent of the award is based on observed teacher performance as evaluated against the *TAP Teaching Skills, Knowledge, and Responsibility Standards*; 30 percent on a value-added measure of classroom-level achievement gains; and 20 percent on a value-added measure of school-level achievement gains. The TAP foundation recommends setting aside $2,500 per teacher for annual performance awards. As an example of a possible range in performance bonuses, teachers at one middle school in South Carolina received performance awards varying from $400 to $4,300 during the 2004-05 school year.

How Schools Become TAP Schools. Selection as a TAP school occurs via a competitive process. Typically, a state department of education or district superintendent invites schools to learn about TAP and apply for the program. TAP requires that teachers from prospective TAP schools vote to express support for the program. Candidate TAP schools also need to show an ability to provide financial support for the program. Ultimately, selection as a TAP school depends on the ability of the schools to implement, fund, and sustain the program, as well as on demonstrated faculty support. The TAP Foundation also seeks to expand where opportunities already exist, often by conducting district pilots or by working with clusters of schools.

Operational Dates. Since the inception of the program in 2000, the number of TAP schools has changed each year, with some schools starting TAP and others ending their participation in the program. The main reason for schools discontinuing TAP appears to be a lack of funds sufficient to maintain the teacher incentives. Changes in administrative support also may lead to the termination of the program. In one case, for example, the superintendent that supported the district's participation in TAP left the district, and the program did not continue under his successor.

Policy Relevance. This program has become well known among education policymakers and researchers and has received considerable media coverage during its five-year existence.

The TAP Foundation website also features supportive quotes about TAP from national, state, and local policymakers, including the U.S. Secretary of Education, senators, governors, superintendents, and teachers union officials. Interest in the program likely stems in part from its distinctiveness as one of the only private initiatives of its kind. In addition, the spread of TAP from five schools in one state in 2000 to over 100 schools in multiple states in 2005 has helped TAP garner the attention of policymakers as a replicable program. The program also appears to have spurred legislative initiatives, such as Minnesota's Quality Compensation for Teachers (Q-Comp) reform package, which is based largely on TAP principles.

Prior and Proposed Research. Two previous studies analyzed the impacts of TAP on student achievement and teacher attitudes and satisfaction during the initial years of implementation. Most of the authors of the two studies are currently or formerly affiliated with the Milken Family Foundation or its offshoot, the TAP Foundation. The earlier work by Schacter et al. (2002) investigated the impact of TAP during the first two years of its implementation in Arizona. The researchers matched TAP schools with comparable schools within the state and used a statistical model to compare the achievement of students in TAP schools to the achievement of students in the matched comparison schools. A more recent study by Schacter et al. (2004) conducted a similar analysis after three years of implementation in Arizona and one year in South Carolina. Both studies found that the majority of TAP schools posted significantly greater student achievement gains than their matched comparison schools. The studies also found that the majority of teachers in TAP schools supported most aspects of TAP. Support for the performance pay element tended to be low; however, the authors noted that the lack of endorsement for this principle did not appear to diminish the sense of collegiality and teamwork among teachers.

As suggested by the authors, small sample sizes limited the extent to which valid conclusions about the efficacy of TAP could be drawn from these analyses. The earlier study analyzed four TAP schools in a single state, while the more recent study analyzed eleven TAP schools in two states. If we were to study this program, our work would expand the previous analyses by utilizing more TAP schools in additional states, as well as more years of data, including the years before, during, and after implementation.

In addition to this prior research, the TAP Foundation has received a proposal for future research on TAP by Jay Greene of the University of Arkansas. As we understand his proposed research, Greene would focus primarily on the effects of TAP on teamwork and other teacher outcomes. Our work would focus on the impacts of TAP by analyzing student achievement and other outcomes using the data from multiple years and states that are now available.

b. Overview of Possible Study Design

Research Questions. We would attempt to answer the following three questions about TAP with the proposed analysis:

1. Does TAP make teachers perform better in terms of value added to student achievement than they would without the program?

2. Does TAP help schools *attract* teachers who are better in terms of value added and qualifications than the teachers they would attract in the absence of the program?

3. Does TAP help schools *retain* qualified teachers at higher rates than they would in the absence of the program?

Identifying Program Effects. The TAP analysis would exploit variation in the outcomes of interest (test score gains, teacher recruitment, and teacher retention) over time and between schools. Comparison schools will be those that are not implementing TAP but are in the same states as TAP schools.[3] Non-participating schools are acceptable comparisons in this case because the program cannot serve all possible schools. Therefore, the reasons that a school adopts or does not adopt TAP may depend on factors that are random with respect to the outcomes of interest, such as whether the district superintendent or school principal is personally acquainted with the TAP state coordinator or other TAP Foundation staff members.[4]

To further strengthen the comparison group design, we will use all schools of the same grade configuration for which data are available and carefully control for existing variation in district- and school-level characteristics, including those that are based on averages of student characteristics.[5] We will also conduct sensitivity analysis to exclude TAP schools that explicitly applied to but did not implement the program, because including such schools may introduce selection bias. The selection bias would be a concern if one believed that non-participating applicants to TAP were likely to have substantially worse (or better) outcomes than TAP schools would have had in the absence of the program.

Variables and Measurement. The three main outcomes of interest are teachers' productivity (more specifically, value added to student achievement growth), retention, and recruitment. We will focus primarily on productivity, since all of the states we have identified for inclusion in an analysis of TAP schools would have test score data available. We will estimate value added by using regression-adjusted individual level test score gains where possible. Otherwise, we will use test score gains that are adjusted using school-level covariates. The gains in the case of school level data would have to be based on scores for a given grade/school in year *t* and the same school/previous grade in year *t*-1. If measures of teacher quality or qualifications can be constructed, we will compare teacher qualifications of TAP with

[3] In states such as Florida, Louisiana, and South Carolina where each district typically contains a large number of schools, non-participating schools in the same *district* as TAP schools may serve as a comparison group.

[4] We considered using as a comparison group the prospective TAP schools that failed to adopt the program due to a vote of support from faculty members falling just short of the threshold required for adoption (typically 70 percent). However, we learned that, except in the case of Minneapolis, the reasons a prospective school might not adopt are rarely related to buy-in from faculty and more often have to do with the availability of outside funding to support the teacher incentive payments. In Minneapolis, we do have the option of faculty vote percentages as an index for a regression discontinuity design.

[5] We will use both matching (e.g., propensity score matching) and covariate adjustment and test the sensitivity of the results to different approaches, such as the matching algorithm and the functional form of the covariate adjustment.

those of comparison schools to estimate the program's effect on the composition of the teacher workforce, an effect that can be the combination of recruitment and retention effects. To disentangle the recruitment and retention effects, we will examine transfer rates and turnover rates where possible. We believe that such data can be obtained from at least some of the states (Florida, Arkansas, and Colorado) and are exploring whether they can be obtained or approximated for the other states.

Data Availability. One challenge in studying TAP is that its schools are located in many states, which may vary widely in their testing and data collection. We have identified seven states (shown in Table 2) in which at least some public schools were using TAP by the 2004-05 school year. We believe it would be feasible and cost-effective to obtain data (on student test scores, schools, and, in some cases, teachers) from all seven states. We would conduct separate analyses by state and combine the results by reporting the seven state-specific estimates (if all seven states were used), as well as the average and range. One approach to doing this would be to construct a formal meta-analysis, which would allow us to maximize the statistical information about the estimated effects by state and their standard errors.

TABLE 2

LOCATION OF TAP SCHOOLS BY STATE

State	Number of Schools That Ever Participated in TAP, Through 2004-05 School Year	Number of Schools That Started TAP and Later Discontinued Program, Through 2004-05
Florida	22	19
Colorado	15	0
Arkansas	14	0
South Carolina	11	4
Minnesota (Minneapolis)	3[a]	0
Arizona	6[b]	3
Louisiana	6	0
Total	82	30

[a]Seven Minnesota schools began operating TAP during the 2004-05 school year. As discussed in the text, we propose focusing the analysis of TAP in Minnesota on the three TAP schools located in Minneapolis in order to exploit the rich data available from the Minneapolis school district.

[b]Seven Arizona schools participated in TAP. However, we will not include in the analysis one TAP school that served only Kindergarten through the second grade due to a lack of testing information for early grades.

The state of Florida has especially comprehensive data available. The Florida Department of Education operates a K-20 Education Data Warehouse from which extensive data on any student or teacher in the state can be extracted for the years going back to 1998-99 and linked over time for teachers or students by using unique ID codes. The Florida Comprehensive Achievement Test (FCAT) has been administered in grades 3 to 8 since 1998 and in grades 3 to

10 since 2001. In addition, the data warehouse has information on students (free lunch program status, enrollment information, race, gender, disability status, English proficiency, course information, and college-going) and teachers (average years of experience, other courses taught, and limited information on certification by subject area). The data would allow us to estimate the value added by teachers before, during, and after implementation of TAP, accounting for the schools that discontinued the program and the teachers who may have transferred into or out of TAP schools.

There are two sources of data for Colorado. The state has data available at the aggregate (school) level for grades 3 to 8 dating back to 2000-01 that can be used to study its TAP schools before and during implementation of the program. In addition, all 15 TAP schools and over 300 non-TAP schools use the tests developed by the Northwest Evaluation Association (NWEA), whose scores can be accessed as part of NWEA's Growth Research Database. NWEA data are available prior to program implementation for 10 of the 15 Colorado TAP schools and during program implementation for all 15 schools, providing a source of individual student-level data for the state's analysis.

Within the state of Minnesota, the Minneapolis Public School District maintains data particularly well suited for an analysis of TAP. The Minneapolis district has administered the Northwest Achievement Levels Test (NALT) to grades 2 to 7 and 9 since 1998-99. Based on our conversations with the public school research director, we believe we can link NALT test score data by grade to data on teacher teams for each school in the district. One strategy for exploiting this information would be to focus the TAP analysis on Minneapolis, comparing TAP schools to non-TAP schools within the district.

Data from Arizona, Louisiana, South Carolina, and Arkansas can be obtained from those states' websites. In Arizona, test scores are available by school for each of grades 2 to 9 from 1998-99 to 2004-05, allowing us to observe student achievement prior to and during TAP implementation; we can also observe post-TAP scores for three of the schools that discontinued the program. Louisiana test score data for grades 3 to 9 dating back to 1999-2000 enable us to compare data on TAP and non-TAP schools before and during implementation for the six Louisiana schools that began the program before the 2005-06 school year. For South Carolina, school-level test scores for grades 3 to 8 available from the state website predate TAP implementation in the state and continue to the present.[6] Data availability is more limited for Arkansas. While school-level testing data are available for grades 4, 6, and 8 starting in 2000-01, the most recent year of data presently available is for the 2003-04 school year, thus allowing for only one or two years of observations during TAP operation.

c. Summary Comments

Demand for research evidence on the effects of TAP will be very high, so any new data analysis generated by independent researchers not affiliated with the TAP Foundation would be

[6] Although NWEA data is available for some TAP schools in South Carolina, we do not plan to use the NWEA database for this state due to inconsistent availability of the NWEA tests across South Carolina schools.

valuable. The program is well defined and replicable, and it demonstrates a substantial shift in how teachers are traditionally compensated.

The presence of TAP in multiple states presents both a challenge and an opportunity. It will be challenging to obtain data from multiple sources and combine findings from different contexts. At the same time, however, findings from a multistate study would have wider applicability than a study that was too dependent on the context of one district or state.

For a successful study of TAP, we would have to seek cooperation and acceptance from the program operators while maintaining the independence of the research effort. We have begun initial conversations with TAP Foundation staff and feel confident that a study of the program would feasible and informative.

2. CHARLOTTE-MECKLENBURG (NC) PAY FOR PERFORMANCE PILOT

a. Overview of Program

Goal. The goal of the Charlotte-Mecklenburg pay for performance program is to improve student achievement in low-performing schools by rewarding staff based on their attendance, professional development, and student performance.[7]

Program Structure and Operation. The Pay for Performance pilot program provides cash bonuses to staff in the Charlotte-Mecklenburg school district for attaining certain goals. All staff in the pilot schools are eligible to participate in the program. In the first year (2004-05), the bonus was contingent upon staff meeting individualized goals for student achievement. For teachers, the goals were based on growth or improvement in student test scores, such as North Carolina's End of Grade or End of Course tests, as well as local tests. Other staff were given goals related to student outcomes; for example, social workers had to decrease their students' drop-out rates. If they met their achievement goals, staff could earn an additional bonus based on their own attendance and participation in professional development.

In the second year of the pilot, the student achievement goals and attendance/professional development measures were separated. Only teachers were given student achievement goals, although these teachers could earn the attendance/professional development bonus even if they did not meet their student achievement goals. The non-instructional staff could earn a bonus based on attendance and professional development.

Eligibility. At the elementary and middle school level, Charlotte-Mecklenburg has 48 low-performing schools, designated as "FOCUS" schools, 6 of which (3 elementary schools and 3 middle schools) were selected at random to participate in the pilot by the Center for Policy Research in Education (CPRE). Of the 42 non-pilot schools, 29 are elementary schools and 13 are middle schools. In addition, there are six FOCUS high schools, five of which implemented pay for performance, but these schools were assigned purposively and three of them implemented their own version of the pay for performance program. Therefore, we propose to focus the study on just elementary and middle level schools.

The summer before the pilot began, all district teachers were informed of the program, although the specific criteria for earning the bonus were not revealed until September or October. During that summer, teachers in non-FOCUS schools were offered a one-week transfer period to fill vacancies in the pilot schools. Approximately 12 teachers transferred during this period.[8]

[7] Charlotte-Mecklenburg school district offers several teacher incentive programs. One of those is a district-wide signing bonus program begun in 2000, which we considered as another possible study opportunity for this report. However, we determined that a convincing quasi-experimental analysis would not be possible, since the program was implemented district-wide in a district that is fairly unique in the state. Therefore, no feasible comparison group could be constructed to approximate the counterfactual outcomes.

[8] These transfer teachers may have differed from other FOCUS teachers, particularly if they transferred in hope of obtaining the bonus. If the teachers who transferred to pilot schools were above average in quality, this would represent a recruitment effect of the program. However, the small number of transfer teachers suggests that any positive impacts of the pilot program on teacher performance (value added to student achievement) are more likely the result of productivity effects.

During the second year, there was not a transfer period, and the teachers were notified about changes to the pilot and the criteria to earn the bonuses by early September.

Within the pilot schools, participation of the staff was voluntary and did not affect eligibility for or receipt of other bonuses offered by the district or state. Participation was not automatic; staff had to agree to be part of the program on a year-to-year basis. Approximately 85 percent of eligible teachers agreed to participate during the first year. Data from the second year are not yet available.

Award. In the first year, the bonus was centered on student achievement. Although all staff had student achievement goals, our focus here is on classroom teachers, who are likely to have a more direct impact on student achievement than other staff, such as administrators or janitors. Teachers who volunteered for the program were given individualized goals for student test scores. These goals were formulated by a team typically comprised of human resources personnel, the school principal, and department heads. For the most part, teachers were not part of the teams, although they were solicited for feedback and suggestions. For teachers with classes associated with End of Grade (EOG) or End of Course (EOC) tests, such as Algebra and Biology, goals coincided with the statewide measures of high growth. For teachers to obtain the bonus, an average of their students' scores had to meet or exceed the high growth measures set by the state. The state goals are based on historical statewide averages, adjusted for a cohort's ability level and expected regression to the mean. Other teachers, who led classes that did not have EOC or EOG tests, were given a variety of goals, such as increasing the percentage of students who passed a local test.

In the first year of the pilot, teachers could earn an additional bonus based on attendance and professional development if they met their achievement goals. This bonus was given if a teacher missed four or fewer days in a school year and attended at least thirty hours of professional development. In the second year, the attendance/professional development bonus was not based on attaining achievement goals.

Teachers in the pilot schools were awarded $1,400 for attaining the achievement goal and $600 for meeting the attendance/professional development goals. For the first year, approximately 200 certified teachers received the bonus, amounting to about 25 percent of the teachers who participated in the program.[9] Roughly half of these teachers received only the bonus for student achievement. The second year of operation is still in progress.

Operational Dates. The pilot was developed in the summer of 2004 and implemented in the 2004-05 school year. The district modified the program for the second year, 2005-06, although the changes mainly affected non-teachers. For example, the major changes were to separate eligibility for the student achievement and attendance/professional development goals, restrict the achievement measures to teachers only, and eliminate the higher bonus for the three

[9] These figures include high schools.

high schools. For the proposed analysis, we would be able to examine outcomes for the years leading up to implementation of the program and the first two program years.[10]

Policy Relevance. Current interest in linking teacher compensation to student achievement is very high; this program addresses the topic directly. In this pilot, teachers are rewarded for their particular students' achievement, based on state test scores. The pilot was designed to use what the district considered objective measures of teachers' efficacy, rather than evaluations (e.g., by the principal or other teachers) that may be more subjective.

In addition, the bonuses for this pilot are similar to amounts offered by other districts' incentive programs, such as signing or retention bonuses. For this reason, bonuses for this program are not prohibitively expensive, and the program could be appealing to many other districts interested in rewarding teachers' performance.

Prior Research. We have not identified any evaluations of this program, nor have the staff in human resources at the Charlotte-Mecklenburg school district learned of any evaluations. The district, however, is known as an innovator in accountability programs and has received media and scholarly attention for other programs (cf, Snipes et al. 2002). For instance, Charlotte-Mecklenburg was one of the first districts in the country to implement a school-based performance award program, which became the model for North Carolina's statewide accountability plan (Johnson et al. 1999).

b. Overview of Possible Study Design

Research Questions. A study of this program would address the question: Did the potential to earn a bonus encourage teachers in eligible schools to raise their students' test scores more than would have been expected in the absence of the program?

Identification of Program Effects. We propose to estimate impacts of the pay for performance program by exploiting variation in student achievement gains and other outcomes across schools, classrooms, and time periods. The basic approach is a comparison school design, but set in the context of a multilevel model where we use all the information contained in the student and classroom level data for the period during and before the implementation of the program. Until we learn more about the assignment process that produced the set of pilot and non-pilot schools, which was reported to be random, we propose two specific approaches to constructing the comparison group, one that assumes assignment was random and one that does not.

The first approach is to assume the assignment of schools to pilot status was truly random and use non-pilot FOCUS schools in Charlotte-Mecklenburg as a comparison group. We would exclude high schools from this analysis entirely because their assignment was non-random. If assignment of the remaining schools is random, this approach is preferable, as it removes any

[10] Continuation of the program beyond the 2005-06 year is uncertain. Charlotte-Mecklenburg may be redesigning its incentives programs to include fewer programs, each with larger incentives.

systematic differences in both observed and unobserved characteristics of pilot and non-pilot schools.

The second approach would use matching, by selecting schools in Charlotte and possibly throughout the state that are similar to the pilot schools on observed characteristics. We would implement this design if we learned that assignments to pilot status were not truly random. There are numerous variables available to researchers for matching, including school enrollment, pupil-teacher ratio, percent of students eligible to receive free/reduced-priced lunch, percent of minority students, and school achievement as measured by annual yearly progress or the annual performance index. By matching schools, we would reduce the likelihood that other factors, such as relative school advantage, might lead to observed differences between teachers.

Because of data availability, we will focus on teachers whose students take EOG and EOC, tests described below. We would compare the achievement of the students with pilot school teachers to those of the non-pilot school teachers. This likely would be a model regressing the 2004-05 EOG or EOC on a host of school, teacher, and student characteristics. School variables would include whether the school is a pilot, and grade levels (i.e., elementary, middle, high school). In terms of teacher characteristics, we could control for years of experience, years of tenure at the current school, licensure, and education. Student characteristics would include age, race, sex, free/reduced lunch eligibility, learning disabled status, and exceptionality status. We also would include a pretest score for the students, such as the EOG in a previous year, or results for a different EOC to control for heterogeneity in the skills and knowledge among students.

Data Availability. The North Carolina Education Research Data Center has statewide information dating back to 1995. Based on our conversations with staff at the Data Center, we believe there would be about a one-year lag from the end of the most current school year for acquiring data that links student scores to teachers. The Data Center has data at many levels, including district, school, classroom, teacher, and student. Student records can be connected from year to year for longitudinal analysis. In addition, students can be linked to the teacher who administered the EOG or EOC test. This is often, but not always, the students' instructor. Since we are unable to determine if a teacher is the instructor of a class, we will use the average characteristics of teachers in a school, rather than characteristics of teachers in any given classroom wherever we believe the teacher-student link is inaccurate.

The Data Center has information on all of the aforementioned variables, such as pupil-teacher ratio, teacher experience, and race and gender of the student. The available student achievement measures are the EOG tests, which are given annually from grades three through eight, and EOC tests (Algebra I & II, Geometry, Biology, Chemistry, Physics, Physical Science, English I & II, Economic-Legal-Political Systems, U.S. History), which can be administered in different grades.

Obtaining data from the Data Center requires an application procedure, Institutional Review Board approval for the project, and sponsorship by a government or nonprofit agency. The Data Center would charge fees based on the number of days of effort required to prepare the dataset for the research project.

c. Summary Comments

Studying the Charlotte-Mecklenburg program would have several advantages. First, the program directly rewards teachers for their students' performance, as well as teachers' attendance and professional development. Second, we are able to select comparison schools from within the same district and so are able to control for all district-level characteristics. Third, the data on teachers in North Carolina are extensive and detailed, going back for more than a decade. Finally, most of the pilot schools were randomly selected, which reduces the possibility that pilot schools are systematically different from non-pilot schools.

A study of the Charlotte-Mecklenburg program also would have some limitations. One is that the sample of pilot schools is small, which makes it difficult to determine whether impact estimates reflect true impacts or idiosyncrasies of the particular schools. Another is that the program is not mature. The pilot has been in place for less than two years and the program already has been modified, although findings from the first year of the pilot would still be informative because the changes did not greatly affect the regular subject classroom teachers.

An important caveat for interpreting the Charlotte-Mecklenburg data is that the district also participates in other bonus programs, particularly for EOG and EOC results. North Carolina has a school-based bonus program for schools that meet or exceed their expected growth measures. Charlotte-Mecklenburg has an add-on to this program that rewards schools that meet or exceed expected growth for all subgroups of students. Consequently, the teachers in the comparison group also have monetary incentives to improve their students' achievement, although the association between the teachers and student achievement is not as direct as for the pilot teachers. Therefore, the proposed analysis will estimate the impact of the *difference in incentive effects* between pilot schools and other FOCUS schools.

Weighing the limitations against the potential for contributing to the knowledge base about merit pay, we believe that this program is a promising candidate for secondary data analysis.

3. CALIFORNIA'S CERTIFICATED STAFF PERFORMANCE INCENTIVE AWARD PROGRAM

a. Overview of Program

Goal. The goal of this program was to foster greater standardized test score increases in low-performing schools throughout California.

Program Structure and Operation. This was a statewide program operated by the California Department of Education (CDE). It provided cash bonuses to all certificated staff[11] in low-performing schools that showed the greatest growth in school test scores from one spring to the next, as measured on the Academic Performance Index (API; more details on this below). Advance knowledge of the potential bonus was meant to spur school staff to help students achieve larger gains than they otherwise might have achieved.

The distinctive features of this program for our purposes are:

- ***Large Bonuses.*** The bonuses were as large as $25,000 per staff member and were tied exclusively to test scores. This type of incentive is relatively extreme compared to others that have been offered around the country, a characteristic that makes this a potentially interesting test case.
- ***Short Time Span***. The program was fully implemented in one year and then was cancelled for budgetary reasons before bonuses could be awarded for the following year. In the second year, the incentives were in place, but bonuses never were awarded. This brief existence suggests that the high cost of this program may limit its sustainability, but it also provides a unique research opportunity, as we discuss below.

Eligibility. For a school's staff to be eligible to compete for the award (to be subject to the incentive), the school had to meet the two criteria: (1) be in the lower half of the API distribution based on its baseline score (e.g., a school in the lower half of the API in spring 1999 would be eligible for awards based on growth between then and spring 2000), and (2) have shown test score growth in the prior year (e.g., in 1998-99, continuing our example). School staff should have known where they stood on these criteria by the January following baseline spring testing.

Award. To qualify for a bonus, schools had to meet three additional eligibility criteria: (1) the school had to achieve growth of at least double its target, (2) all numerically significant subgroups (e.g., racial/ethnic minority groups) in the school had to have made 80 percent of the school's API growth target; and (3) the school's students had to have met a minimum test score participation rate—95 percent for elementary and middle schools and 90 percent for high schools. Of the roughly 4,000 schools initially eligible (in the bottom half of the base API score

[11] A "certificated" staff member is any school employee in a position requiring certification and who holds a document issued by the state teacher credentialing agency authorizing service in a state public school, including a credential, emergency permit, or waiver. Thus, the awards were not restricted to classroom teachers; principals, for example, also received bonuses.

distribution) during the first year the program operated, about 1,300 schools (one-third) met the three follow-up criteria.

CDE ranked all of the qualified schools on their API growth scores and then considered the number of FTE certificated staff at each school on this list. For schools encompassing the first 1,000 FTEs, each certificated staff member would receive a bonus of $25,000; at schools encompassing the next 3,500 FTEs, the bonuses were $10,000 per person; and at schools encompassing the next 7,500 FTEs, the bonuses were $5,000 per person.[12] The program thus called for about 12,000 FTE staff (including classroom teachers, other teachers, and administrators) to share a total bonus pool of almost $100 million. During the one year this program was fully implemented, staff at about 300 schools received a bonus (less than one-fourth of those on the qualified list). Schools near the top of the list had achieved gains of 75 points or more on the API.

Calculating Test Score Growth in California. To understand program eligibility and selection for these bonuses, some background on the API may help. The API is a weighted average of students' standardized test scores. The scale ranges from 200 to 1,000; the statewide goal for every school is an API of 800. Each year since spring 1999, for virtually every school in the state, CDE has calculated and published the base API, the school's growth target for the next year, and the observed change over the past year. The growth target for a school is 5 percent of the difference between its base API and the statewide standard goal of 800. For example, a school with base API of 400 has a growth target of 20 (.05*[800-400]). A school with a base API of 700 has a growth target of 5. Scores are measured and targets established for numerically significant subgroups within each school. In most cases, the growth target for each subgroup is 80 percent of the schoolwide target.

Operational Dates. The program was established in 1999 and was on the books for two school years. The first round of awards was based on API growth from spring 1999 to spring 2000. Ultimately, that was the only year for which awards were made, because a state budget crisis led to the program's cancellation. However, funds for the program were not cut until February 2002, well after what would have been the second award year, spring 2000 to spring 2001. Therefore, staff should have entered the 2000-01 school year expecting that the awards would be made based on growth achieved from spring 2000 to spring 2001; those in eligible schools should have had the same motivation to pursue the awards as staff had in the prior year.[13] This suggests that we could study API gains for both years, not just for the one year for which awards were actually distributed, using as comparison points the data from before and after those years.

Policy Relevance. This program attracted fairly substantial national attention during its brief existence, such as several stories in *Education Week*. This attention probably was due largely to the size of the bonuses. None of the other programs we identified in our extensive

[12]Receipt of the full amounts was dependent on agreement by the local union; where unions failed to negotiate with the local school board and agree to this arrangement (such as in Los Angeles Unified), state law called for funds to be distributed on the basis of staff salaries.

[13] A high priority of the implementation analysis will be to confirm this claim through interviews.

search provided bonuses anywhere near as large as the maximum bonus of $25,000 per person offered in California. However, the Houston (Texas) Independent School Board recently approved a teacher performance bonus plan that could be expanded to provide as much as $10,000 in merit pay for teachers.

Also, the California program is one of the few we identified that tied bonuses directly and exclusively to student test score gains. It seems likely that national, state, and local education policymakers would be interested in the results of a study that sheds light on whether a program like this can change teacher behavior and result in greater student achievement gains.

Prior Research. We have not identified any rigorous evaluations of this bonus program, but the program and API scores in general were the subject of much comment and some analysis. The California Budget Project (March 2001) identified several school characteristics associated with API scores and claimed that 80 percent of the variation in scores could be explained by social and economic factors. It also claimed that this particular award program was biased against the lowest performing schools, because they faced the highest growth targets and had the fewest means of meeting them (April 2001). It advocated that awards be based solely on growth scores, and not be restricted to those schools exceeding twice their growth targets.

An article in the *Orange County Register* newspaper (8/13/2002) pointed out that many schools that received the certificated staff bonuses in the first year had very different outcomes (including substantial API losses) the following year, and went on to identify factors that might account for such variability, including school size and student mobility.

UC Irvine economist Justin Tobias published a regression analysis which showed that judging schools based on API growth was biased against high-performing schools (2003). Two California State-Northridge professors examined API scores from 1999 to 2003 and concluded that reward systems should account for differences in baseline scores and school and community characteristics (Driscoll and Halcoussis 2005). The American Institutes for Research also published an extensive analysis of California test score changes, under contract to CDE (O'Day and Bitter 2003). If we were to study this program, our work would be informed by these studies.

b. Overview of Possible Study Design

Research Questions. A study of this program would address the question: Did the potential to earn substantial bonuses encourage teachers in eligible schools to raise their students' test scores more than would have been expected in the absence of the program?

Identification of Program Effects. The effect of the California program on student achievement can be estimated using a panel data regression discontinuity design. This approach would examine the relationship between API growth scores and baseline test scores[14] before, during, and after program implementation, with statistical controls for school and community

[14]API scores were not available prior to spring 1999; we would need to acquire other testing data.

characteristics. In particular, the eligibility criteria, which are continuous and include arbitrary cutoffs, would be used as statistical controls so that the program effect would be identified as a discontinuity in the relationship between the eligibility criteria and the outcome at the arbitrary cutoff that determines eligibility. Schools would serve as the unit of analysis, in part for simplicity and cost-effectiveness, and in part because eligibility and bonuses were determined at the school level. Any net discontinuity at the eligibility point would be interpreted as the incentive effect of program eligibility on the outcome, such as API growth.

Variables and Measurement. The main outcome variable would be API growth scores, the same measures on which schools were ranked for staff bonus consideration when this program was active. However, unlike the program, we will add several independent variables to control for other factors that might have affected API growth scores for eligible and ineligible schools alike. These measures might include indicators of school size and student characteristics.

Data Availability. The CDE website provides links from which to download extensive data for every school and year since the API was introduced in 1999. Each statewide database includes all the data that are part of an individual school's API base or API growth report, including: school, district, and county name; base API, API growth target, and API growth; percent of students tested; number of students included in base and growth; school's decile in API distribution; API information for 100 comparable schools; API information for seven racial/ethnic subgroups and for socioeconomically disadvantaged students; student background characteristics (percentages in seven racial/ethnic groups, percentage in free or reduced-price lunch program, percent of English language learners, student mobility, and parent education levels); average class size; percentage of teachers fully credentialed; total enrollment on testing dates; number of students exempted from testing based on parent request; and, apparently, the weights and number of valid scores on test subcomponents (e.g., English-language arts, math, science).

As resources for interpreting the data and obtaining supplemental data, the MPR team has extensive contacts among California-based researchers; many of whom are familiar with state data that we would need to use and could provide advice and consultation to the project:

- David Rogosa of Stanford, an expert in the measurement of growth in educational indicators, who has prepared several reports on the API
- Justin Tobias of the University of California at Irvine, Department of Economics
- Tom Kane, on leave from UCLA, now at the Harvard Graduate School of Education
- Julian Betts of UC-San Diego's Department of Economics

c. Summary Comments

One potential drawback of studying this program is that we would have to rely on the school-level data described above. We believe that it will not be feasible to obtain individual level student test data from CDE.

Another argument against studying the California program has to do with its policy relevance. The lessons learned from a study of this program may have limited applicability to other programs because the design was so peculiar. Bonuses of $25,000 are highly unusual and, as in California, probably are unsustainable anywhere they would be implemented; certainly the size of the bonuses in California contributed to the program's cancellation before the second year was completed. On the other hand, one could argue that studying such an extreme case would provide the most efficient way to test the hypothesis that cash bonuses do not raise test scores; if bonuses at the upper extreme do not result in substantial improvements, there probably is little need to test for such effects in programs that give smaller rewards, which includes nearly all programs of policy interest.

Despite some limitations, secondary data analysis to study the California program is worth considering for several reasons. The study would be relatively easy and straightforward and, in all likelihood, could be completed quickly, due to the ready availability of data. It would use a strong quasi-experimental design, and it would directly address the link between individual teacher bonuses and student performance, an important policy research issue.

4. CINCINNATI TEACHER EVALUATION AND COMPENSATION SYSTEM

a. Overview of Program

Goal. The goal of this program is to enhance teacher professionalism and boost student achievement by linking teacher pay to the attainment of progressive levels of teacher mastery and performance as measured by classroom observations and reviews of teacher portfolios.

Program Structure and Operation. Cincinnati has replaced the traditional teacher salary structure of automatic advancements based on years of experience and graduate degrees with a system wherein promotions are tied to teachers' evaluations based on 16 criteria (listed in Table 3).[15] The criteria cover four teaching domains: preparing for student learning, creating an environment for learning, teaching for learning, and professionalism.

Evaluation teams review a portfolio of teacher lesson plans and observe classroom practices to determine their ratings of teachers. Annual ratings provide formative guidance and feedback to teachers. "Comprehensive" reviews, which generally take place once every five years, place teachers into one of five mastery levels, which determine their salary range.

The distinctive features of this program for our purposes are:

1. ***Replacement of the Uniform Salary Schedule***. Rather than supplement an existing seniority-based salary schedule with bonuses, the Cincinnati system ties permanent pay increases to movement up a career ladder where advancement is not automatic.
2. ***Rotation of Annual and Comprehensive Reviews***. A comprehensive review of a teacher's performance on all four domains takes place every two to five years, once the teacher has advanced past the Apprentice level. These are "high stakes" reviews that determine a teacher's mastery ranking, and therefore the salary range. "Low stakes" annual reviews are provided in two domains during all years in which a teacher is not subject to a comprehensive review. Annual reviews serve to provide teachers with constructive criticism and to determine proficiency. Teachers must meet proficiency standards in order to receive experience-based pay step advancements within their mastery rankings.
3. ***Performance-Rated Independent of Student Test Scores***. The evaluation of teacher performance is based on the judgments of informed peers regarding the extent to which a teacher is following professional and pedagogical norms thought to contribute to student learning. The test scores of students are not incorporated into the review in any way.

[15] The 16 specific criteria used for the evaluation are linked to Charlotte Danielson's *Framework for Teaching* (Danielson 1996).

TABLE 3

SUMMARY OF THE 16 PERFORMANCE CRITERIA BY DOMAIN

Domain 1: Planning and Preparing for Student Learning

- Incorporate multicultural sensitivity into lesson plans
- Write clear, multidisciplinary instructional objectives focused on high expectations and individual learning needs
- Use assessments aligned with standards and appropriate instructional resources

Domain 2: Creating an Environment for Learning

- Create an atmosphere of universal caring and respect
- Establish a classroom culture of universally high expectations and involvement
- Maintain a safe and disciplined classroom with time focused on learning

Domain 3: Teaching for Learning

- Know the fundamental knowledge and skills students need prior to learning new material
- Communicate effectively
- Practice interactive teaching that promotes participative learning
- Promote conceptual thinking, critical thinking, and real-life applications
- Obtain information on student progress and challenges in a timely and reliable fashion
- Consider individual student cultural backgrounds and needs and seek effective instructional approaches for each student.

Domain 4: Professionalism

- Record student progress towards academic goals and rubrics for grading
- Inform families about the academic and social development of their children
- Be a reliable team player in the school
- Participate in professional development activities

Eligibility. Most teachers in the Cincinnati Public School District are required to participate in this evaluation and compensation system as a condition of their employment; it is the only teacher compensation system for the district. The only exceptions are teachers with 16 or more years of experience as of spring 2001, who were rendered exempt from the system and remain under the traditional salary structure (Milanowksi and Kimball 2003). Also, see below for details on the phase-in period, during which veteran teachers were promoted and compensated under the traditional system.

Teacher Evaluations and Award Determinations. Comprehensive and annual evaluations are conducted by two-person teams comprised of one Teacher Evaluator and one administrator. The Teacher Evaluators are teachers specially trained to conduct the evaluations; the administrators are principals or assistant principals. All evaluators must reach acceptable levels of inter-rater reliability during training.

Comprehensive reviews involve portfolios of teacher work and classroom observations. The portfolio includes sample lesson plans, student work, statistics on attendance and family contacts, and a narrative on professional development activities. For each comprehensive evaluation, the administrator on the team conducts at least one classroom observation, carefully reviews the teacher portfolio, and scores the teacher on the criteria within the domains of "planning and preparing for student learning" and "professionalism." The Teacher Evaluator conducts three or four additional classroom observations and rates the teacher on the criteria related to the domains of "creating an environment for learning" and "teaching for learning."

Promotions are based on mastery levels keyed to ratings that range from 1 (lowest) to 4 (highest) in each of the four domains. The domain scores are averages of individual criterion ratings within each domain. Table 4 presents an overview of the system, including the sequence of mastery levels, the timing of comprehensive reviews, and the potential positive and negative consequences of review results.

Low stakes annual reviews, which are done mainly to provide feedback to teachers, are conducted in every year that a teacher is not subject to a comprehensive review. Annual reviews are limited to evaluations of "teaching for learning" and one other domain selected by the teacher. Annual reviews also determine whether or not a teacher will advance to the next experience-based salary step; teachers deemed "proficient" based on their annual review receive the scheduled pay raise commensurate with their mastery level and years of experience.

While teachers undergo a more routine review annually, the comprehensive reviews that trigger promotions are conducted every few years. All teachers new to the profession are automatically classified as Apprentice prior to their initial review, which occurs at the end of their first year. New teachers at the Apprentice level must attain a Novice ranking by the end of their second year or their employment is terminated. Novice teachers must pass the PRAXIS III licensing test and attain promotion to Career rank by their fifth year as a Novice or their employment is terminated. All teachers with mastery rankings above Novice are subject to a comprehensive review at least every fifth year. Failing to pass to the next career ladder rung beyond Novice means that promotions are delayed.

TABLE 4

CINCINNATI PUBLIC SCHOOLS PERFORMANCE-BASED PAY SCHEDULE, 2002-03

Mastery Level	Requirements	Timing of Comprehensive Evaluation	Consequence of Failing to Advance	Number of Steps	Salary Range
Apprentice	New teacher	First year (and second if needed)	Termination	0	$30,000
Novice	Licensed and 2s in all domains	Within five years of previous	Termination	3	$32,000-35,750
Career	3s in all domains	Within five years of previous	No raise	4	$38,750-49,250
Advanced	4s in teaching for learning and one other domain, 3s in other two domains	Within five years of previous	No raise	3	$52,500-55,000
Accomplished	4s in all domains	Within five years of previous	NA	2	$60,000-62,500

Note: Adapted from Milanowski and Kimball (2003). This is the most recent information available.

Teachers can request an early comprehensive review after reaching the Novice 3 level, which requires a minimum of four years teaching experience in the district. Upon receiving a comprehensive review, Novice level teachers can move up to Advanced or Accomplished, depending on their review scores.

Teachers can move both up and down the career ladder. Teachers whose evaluations would drop them to a lower ranking than previously attained are eligible for a follow-up review in the next year. If the follow-up confirms the "slippage," the teacher is dropped to the lower mastery ranking and his or her salary is reduced accordingly. Veteran teachers dropped to the Novice rank are placed on probation and monitored by administrators.

Voluntary Lead Teacher Program. As part of the effort to enhance teacher performance in the district, Cincinnati has also established a voluntary "Lead Teacher" program to provide teachers with non-classroom, performance-enhancing experience. Teachers who voluntarily apply for and become Lead Teachers spend three years out of the classroom in a non-teaching position such as peer evaluator, educational consultant, mentor, or program facilitator. After three years, Lead Teachers return to the classroom but maintain the annual $5,000 to $6,500 Lead Teacher bonus in addition to their normal salary. The bonus stays with Lead Teachers, as long as they maintain Advanced or Accomplished status as determined by each comprehensive review.

Any teacher who has reached the Novice 3 level or higher can apply to become a Lead Teacher. After applying to the program, teachers undergo an application and interview (Phase 1) in the spring and, if they pass Phase 1, a comprehensive evaluation (Phase 2) in the fall of the following school year. Applicants who achieve Advanced or Accomplished mastery rankings in their comprehensive evaluations become Lead Teachers.

The only exception to the normal evaluation systems is for teachers who are Nationally Board Certified. Nationally Board Certified teachers who apply to become a Lead Teacher can bypass the comprehensive evaluation (Phase 2) and become Lead Teachers immediately after passing the application and interview process (Phase 1). Nationally Board Certified teachers still have to maintain Advanced or Accomplished status in all future five-year comprehensive evaluations to retain Lead Teacher status.

Operational Dates. A pilot program was implemented in 10 schools in 1999-2000. The system was revised based on the pilot experience, and was designed to take effect district-wide in 2002-03. Based on negotiations between District officials and the teachers union, the new performance-pay system was subject to staged implementation from 2002-03 through 2004-05, with only new teachers, teachers without tenure, and volunteers subject to the new performance-based salary system. The program was phased in for the rest of the district starting in 2005-06.

In the first year of the phase-in (2005-06), all 9th year teachers (new hires in 1996-97) will undergo comprehensive reviews. One or more groups of veteran teachers will be phased into the system each successive year until all teachers hired after 1985 have been evaluated and incorporated into the new evaluation system.

Although the phase-in is meant to be systematic, based on a teachers' year of hire, it will also depend on capacity to conduct comprehensive reviews. The district has 20 full-time teacher evaluators and 8 part-time evaluators. Because the number of new hires varies each year, the district may be able to speed up the phase-in process. For example, all 7th and 10th year teachers (new hires in 1998-99 and 1995-96) will undergo comprehensive evaluations during the 2006-07 school year.

Policy Relevance. The Cincinnati performance-pay system has been recognized in the print media as a pioneering and far-reaching effort to improve the quality of teaching. Both the *Cincinnati Post* and the *Cincinnati Enquirer* have run series on the new teacher pay program. *Education Week* has published at least three stories on the program, including an op-ed that questions the very practice of teacher merit pay. The program was described by a New York Times columnist as "a radical experiment in teacher pay [which] could become a national model if successful" (Rothstein 2001).

The conceptual model underlying the reform forecasts that the system-wide implementation of the plan will produce short-term effects of attracting and retaining more high-performing teachers, improving the performance of continuing teachers, and forging a shared consensus regarding quality instruction. These three short-term effects are predicted to result in improved instruction in the medium-term and improved student achievement in the longer-term (Milanowski and Kimball 2003).

Prior Research. The Cincinnati Teacher Evaluation and Compensation System has been the subject of at least three academic working papers. The system was developed in consultation with Allan Odden of the University of Wisconsin-Madison and the Consortium for Policy Research in Education (CPRE). In 1999, Odden and a colleague co-authored a CPRE working paper that described the process of developing and negotiating the design of the pilot program (Kellor and Odden 1999). A subsequent working paper by Milanowski and Kimball (2003) of CPRE updated the parameters and implementation schedule of the Cincinnati program, as well as a similar program in Washoe County, Nevada.

The most sophisticated analysis of the Cincinnati program to date was reported by Milanowski in a separate CPRE working paper in 2003. He used a hierarchical linear model (HLM) to determine the extent to which variance in student test scores that was not explained by student demographics and prior achievement could be explained by variation in teacher performance evaluations. He identified a small-to-moderate association between teacher evaluations and unexplained variance in student test-score performance, concluding that the association established the criterion-related validity of the performance assessment. Because his goal was merely to determine the validity of the evaluation system, not its impact on achievement, he drew no conclusions regarding whether or not system-induced teaching improvements actually caused the test-score variation (Milanowski 2003).

A new study by IES would be an independent assessment (not by the program designers) of the effect of performance incentives on teacher quality and behavior. Specifically, it would shed light on the question of whether or not students benefit academically when the teachers in an entire urban district are subject to a performance-pay system.

b. Overview of Possible Study Design

Research Questions. The primary questions of interest are: "Do performance-pay systems based on peer evaluation of preparation and classroom skills generate student test-score gains?" and, "If so, how long does it take such gains to manifest themselves?"

But another possible hypothesis to be explored is the "incentives versus professionalism" tradeoff. If the comprehensive review is only summative, providing an incentive for teachers to improve, then we would expect the student test scores for the teacher under review to be better for the year of the comprehensive review than the year after the review, when they could relax and wait another five years to go through their next comprehensive evaluation. However, if the review is formative, meaning that it provides feedback from the reviewers about what the teacher is doing well and what she can do to improve her teaching, then her students' test-scores should be higher the year *after* a review, compared with the year of a review. We plan to conduct implementation analysis, relying primarily on interviews, to assess the degree to which the comprehensive evaluations are formative as well as summative and whether they take into account performance for more than just one year.

Identification of Program Effects. We propose to test the effect of Cincinnati's comprehensive evaluation system on teacher effort by observing deviations in the teacher experience profile at the points where teachers are eligible for comprehensive evaluation. As noted above, prior research on the Cincinnati program focused on the validity or fairness of the

system (Milanowski 2003). We aim instead to estimate the program's impact on teacher performance in the classroom as measured by student achievement. This impact can be identified if we believe that the relationship between teacher experience and performance (value added, as measured by adjusted gains in student achievement) is continuous. If teachers increase their effort in response to the high stakes evaluation, value-added indicators should be higher in years during which the teacher has a comprehensive evaluation, all other things being constant.

One drawback of this approach is that it only captures one aspect of the behavioral response to this type of teacher pay system, the role of comprehensive evaluation relative to routine annual evaluations. To the extent that the program leads to permanent increases in teacher quality, or that it changes the culture of teaching for all teachers, we will not be able to measure the full impact. Another potential drawback, although more subtle, is the idea that teachers in Cincinnati are able to influence the timing of their comprehensive evaluations. For example, if a teacher had information that a principal she felt would rate her performance highly was about to leave the school in the following year, she might seek an early comprehensive review. By getting a high score on the comprehensive review and a lower score in the annual evaluation under the new and potentially less sympathetic principal, the teacher in this example would appear to be more productive in the comprehensive review year purely as a result of who rated her.

A second source of identification of the incentive effects of the system is the difference in performance between teachers with 16 years of experience or more, the level used to "grandfather" existing teachers into the old system, and those whose experience level was just under 16 years at the time of the policy change. Because the cutoff in the experience level at which teachers are no longer subject to the Cincinnati program is arbitrary, we would expect differences in performance between teachers of such similar experience levels to be due largely to the program. To the degree that the phase-in period instituted a grandfather effect for teachers at other experience levels, we can extend this design to include veteran teachers at a variety of experience levels in different years of the observation period.

Variables and Measurement. The primary outcome variables would be student test scores in reading and mathematics. Other data would include student demographics (ethnicity, gender, receipt of free/reduced price lunch, special education status) and teacher characteristics, including when they were evaluated and their scores, years of experience teaching in this district, and grade level taught.

Data Availability. We propose to obtain data directly from the Cincinnati School District, as did Milanowski et al. (2003). The dimensions of the analytic database will be limited by a number of important conditions. First, sequential test-score results will be available only for students in grades four through eight who have been enrolled in the District for two years. Approximately 14,600 of the 38,000 students in the District are likely to be within the tested grade range, and about 10,000 of them are likely to have sequential test-scores.[16] More important, the District reports that it employs 3,200 teachers.[17] Assuming that this number

[16] These calculations assume that 38.5 percent of enrolled students are in grades four through eight and 70 percent of those students participated in the accountability testing in the most recent two years.

[17] We are awaiting more information on who is included in the definition of 3,200 "teachers."

includes teachers acting in administrative and specialist roles, we might expect that about two-thirds—or 2,100—are classroom teachers with at least two years of experience. Of this 2,100, we can assume that only one-fifth of them—or 420—underwent a comprehensive review in one of the previous two academic years. Of these teachers, about 162 of them (38.5 percent) likely taught in grade levels four through eight for which sequential student test-scores would be available. Thus, assuming approximately 15 fully tested students per classroom, the dimensions of our sample are likely to be the 2,400 students in the classrooms of the 162 teachers for whom the study conditions apply. Unfortunately, we do not have estimates at this time of the number of teachers that could contribute the regression discontinuity analysis based on years of experience.

c. Summary Comments

Cincinnati's teacher pay system provides a good example of a program that relies on formal evaluation of teachers across a spectrum of characteristics associated with "good teaching"—moving beyond a simple examination of student test scores—and ties these ratings to both teacher advancement and compensation, with the possibility of dismissal for poor performers.

Several approaches offer the possibility of identifying behavioral parameters of interest to policymakers, given the structure of the program and District. The District has a history of providing individual-level data to evaluators, and the database that would inform the analysis appears to be adequate to identify program-induced effects if they exist. As such, the Cincinnati Teacher Evaluation System is a promising candidate for further study.

5. MISSOURI CAREER LADDER PROGRAM

a. Overview of Program

Goal. The goal of Missouri's Career Ladder program is to improve student achievement by offering teachers opportunities to earn extra pay for extra work and professional development, with eligibility for these opportunities being a function of their observed performance and portfolio of work. Policymakers hope that the incentives created by the availability of such opportunities as well as the activities themselves improve academic services, programs, and learning outcomes for students.

Program Structure and Operation. Through the Career Ladder program, teachers who meet statewide and district-level performance criteria are eligible to receive supplementary pay for Career Ladder responsibilities, which can be extra work or participation in professional development activities. The program does not replace the regular salary schedule. Career Ladder responsibilities must be academic in nature and directly related to the improvement of programs and services for students. The Career Ladder has three stages, based on years of experience and other factors. To move up the ladder, teachers are assessed at each stage through periodic observations and evaluations of documentation. Each successive stage offers the opportunity to receive more supplementary pay for Career Ladder responsibilities: up to $1,500 for Stage I, $3,000 for Stage II, and $5,000 for Stage III. Out of more than 65,000 teachers in 524 districts statewide, more than 17,000 teachers (26 percent) from 333 districts (64 percent) are participating in the Career Ladder program during the 2005-06 school year.

The Missouri program is distinctive among the various teacher compensation reforms we have examined because it is the most mature. The program has been in existence since 1985, outlasting dozens of programs that were first introduced around the country at the same time. Also, the Missouri Career Ladder is unusual in how it mixes teacher performance, tenure, and extra responsibilities to define salary supplements. Teachers must advance along the Career Ladder based on tenure and progress in performance as rated by classroom observers, yet the bonuses actually are given for extra responsibilities. The Career Ladder advancement accounts for only the amount of extra responsibility and the rate at which the extra work is compensated.

District Participation. Missouri's program operates statewide, and districts must choose to participate and provide matching funds. Districts wishing to implement a Career Ladder program must submit a District Career Ladder Plan (DCLP) to the Missouri Department of Elementary and Secondary Education (DESE). DESE approves plans that meet state guidelines for improving academic services and programs for students. DCLPs must be aligned with a statewide Missouri School Improvement Program. They also must include curriculum development plans, professional development plans for teachers, guidelines for teachers' Career Development Plans (CDP), and an instrument for Performance-Based Teacher Evaluation (PBTE). Once approved, the district need only submit another DCLP if it wants to change its plan.

Career Ladder programs are funded jointly by DESE and participating districts. Poorer districts receive a higher percentage of matching funds from the state. Annually, the state ranks districts according to their per-capita income, and covers 40 percent of the program costs for those in the top quartile, 50 percent for those in the next quartile, and 60 percent for those in the

bottom half. Some districts may not participate in the program because they are unable to afford their share of program costs despite the graduated matching rate.

Teacher Eligibility and Qualifications for Award. To be eligible for the supplementary pay, teachers in participating districts must be serving on not less than a regular-length full-time contract, and must have Missouri teacher certification; they also must formally enroll in the Career Ladder program.

To enroll in the Career Ladder and qualify for awards, teachers must develop a Career Development Plan (CDP) associating each Career Ladder responsibility with either a designated plan or some other instructional improvement. The district Career Ladder Review Committee, which is made up of educators (selected by teachers) and administrators, must then approve the teacher's CDP. Through scheduled and unscheduled observations, as well as reviews of their CDP and other documentation such as lesson plans, the teacher is expected to show evidence of performance at or above the expected level on 20 criteria on the district's PBTE evaluation instrument. The criteria span the following six areas: (1) engaging students in class, (2) correctly assessing students, (3) exhibiting content knowledge, (4) professionalism in the school, (5) participation in professional development, and (6) adherence to the district's education mission. There are also specific qualification criteria for each stage of the Career Ladder:

- ***Stage I.*** To qualify for Stage I, a teacher must have five years of teaching experience in the Missouri public schools system and have performed at the "expected" level or above on all criteria on the most recent final evaluation instrument of the PBTE.
- ***Stage II***. To qualify for Stage II, a teacher must have completed two years of service at Stage I of the Career Ladder. The district may waive one year of service at the previous stage if the teacher has spent a total of seven years teaching in Missouri's public schools. The teacher also must have performed at the "expected" level or above on all criteria, and above the expected level on at least 10 percent of the criteria on the most recent final evaluation instrument of the PBTE.
- ***Stage III.*** To qualify for Stage III, a teacher must have completed three years of service at Stage II of the Career Ladder. The district may waive two years of service at the previous stage if the teacher has spent a total of 10 years teaching in Missouri's public schools. The teacher also must have performed at the "expected" level or above on all criteria, and above the expected level on at least 15 percent of the criteria on the most recent final evaluation instrument of the PBTE.

Awards. To receive a salary supplement, teachers must spend a specified amount of time on a certain number of responsibilities outside of their contracted time. Examples of the extra responsibilities that Career Ladder teachers undertake include:

- Extra work—providing students with opportunities for enhanced learning experiences, remedial assistance, and various extended day/year activities

- Professional development activities—participating in professional growth activities, including college classes, workshops, and professional organizations[18]

The district's Career Ladder Review Committee evaluates the teachers to determine if they have carried out their responsibilities and should receive supplementary pay. Almost all Career Ladder teachers receive their supplementary pay. The minimum amount of time teachers are required to spend on these responsibilities is determined by their stage on the Career Ladder, as follows:

- Stage I teachers must spend a total of at least 60 hours on at least two responsibilities
- Stage II teachers must spend a total of at least 90 hours on at least three responsibilities
- Stage III teachers must spend a total of at least 120 hours on at least four responsibilities.

For the 2004-05 school year, the average number of hours spent by Stage I participants was 77, Stage II teachers averaged 110 hours of extra service, and Stage III teacher averaged 141 hours.[19]

Operational Dates. DESE established the program in 1985, and began implementation in the 1986-87 school year, with 2,400 teachers from 63 districts participating. DESE's records indicate that a handful of districts have cycled in and out of participation. For example, whereas 338 districts participated in the 2002-03 school year, 328 districts participated in the 2004-05 school year. This variation over time in district participation status may prove to be useful for identifying program effects, as we discuss below. There were 16,919 teachers participating during the 2004-05 school year. Of these teachers, 3,498 were at Stage I; 4,313 were at Stage II; and 9,108 were at Stage III.

Policy Relevance. This program is relevant due to its Career Ladder structure, the modest size of the supplementary pay, and its longevity compared to other programs. It is one of a number of Career Ladder programs that were established across the nation in response to the recommendations of the influential report *A Nation at Risk* (National Commission on Excellence in Education 1983) to provide supplementary pay to teachers as they moved up the rungs of a ladder. Also, due to the modest size of the supplementary pay, as compared to other programs, a study of this program could provide some ideas about the thresholds at which incentives can have effects. Additionally, although this program has not received as much media attention as others described in this report, it has outlasted similar programs attempted in other states. The

[18] DESE recommends that teachers should not spend more than one-third of Career Ladder hours on college classes and workshops.

[19] These hours approximately translate to supplementary pay at $19.48, $27.27, and $35.46 per hour, respectively for Stages I, II, and III.

number of districts and teachers participating in the program steadily increased after implementation began in 1986, and has leveled off since 2003. Almost all teachers in participating districts currently are enrolled in the program. These features of the program suggest that the program has matured, which allows for easier identification of systemic differences in districts' implementation of the program. For other programs that are not mature, implementation issues may confound the interpretation of study findings.

Prior Research. We have not identified any evaluations of this program. However, there have been some studies of teacher quality in the state, and of teacher recruitment and retention, which could provide some background for any study that we might conduct. Podgursky et al. (2002) used administrative data and found that, although higher-ability teachers (based on their ACT scores) left teaching for other occupations, there was little evidence that they left for higher pay. DESE (2001) reported that, since 1995, there has been a dramatic rise in the percentage of teachers in the state who leave public schools within five years. A report from Southwest Missouri State University (Hough 2000) also noted a continuing trend toward teacher shortages in the state. These reports suggest the need for further analyses of the teacher labor market in Missouri.

b. Overview of Possible Study Design

Research Questions. A study of this program would address the central question: Did the potential to earn supplementary pay for the extra work lead teachers in participating districts to raise their students' test scores more than they would have otherwise?

Some related questions include: (1) Is the extra work of teachers correlated with improved student performance? (2) Does offering supplementary pay increase the amount of time that teachers spend on extra responsibilities more than would have been expected in the absence of the program? Answers to these questions would help explain any impacts identified in response to the central question listed above.

Identification of Program Effects. We propose to identify program effects using both cross-sectional variation, such as *matched comparison* of districts, and variation over time, where we may observe some districts whose participation status, under some circumstances, changes arbitrarily from year to year. For selected participating districts and non-participating districts, we would compare the types and amount of time teachers spent on extra work (intermediate outcomes) if available, as well as test-score gains (final outcome). Controlling for teacher, student, and district characteristics, we will attribute differences in these outcomes to the net effect of the Career Ladder program on teacher performance.

Variables and Measurement. The two main outcome variables at the district level would be (1) mean test scores, and (2) average time spent on extra responsibilities. We will control for other factors that might affect test scores for participating and non-participating districts. These measures include indicators of teacher, student, school, and district characteristics.

Availability of Aggregated Data. We can readily download district-level data from DESE's website (http://dese.mo.gov/schooldata/ftpdata.html).

Test Scores. We will use the proficiency test scores from the Missouri Assessment Program (MAP), which are administered during the spring for Mathematics in grades 4, 8, and 10; and Communication Arts (Reading) in grades 3, 7, and 11.[20]

Career Ladder Hours and Responsibilities. DESE prepares annual reports with the aggregate number of hours and the types of activities that Career Ladder teachers performed in each district. The reports contain the following relevant details by district: (1) number of teachers at each stage, (2) total number of hours spent for each stage, (3) total number of hours spent for each of seven categories of responsibility,[21] (4) average number of hours spent for each stage.

We also can obtain the following data:

- Student characteristics—gender, race, attendance, promotion, suspensions and disciplinary actions, Limited English Proficiency (LEP), economic disadvantage, and disability status
- School/Classroom characteristics—average school/classroom size
- District characteristics—type of district (urban, suburban, rural), size of district, average district household income, per-pupil expenditure

Availability of Disaggregated Data. We also can obtain data disaggregated at the building or individual levels. To obtain such data, we would submit a formal request to DESE, which estimates two weeks between the request and when they provide access to the data, either on a CD or via electronic access for extraction in SAS. Their data are housed at the University of Missouri, Columbia. DESE has the individual-level records available in its Teacher Certification System (educator-level data), the School Core Data System (educator-, course-level data), and MAP Data System (student-level data). We would obtain teacher characteristics, such as years of experience (overall, in Missouri, in district), level of education (highest degree), race, and gender. There is data for individuals with Social Security numbers. The 2002 study by Podgursky et al. indicated that teacher and student identifier codes permitted longitudinal matching of records.

There is data, dating back to the 2001-02 school year, about educators, the number of hours of professional development, courses and assignments, enrollment, and membership. DESE's website indicates the availability of data for teacher certification and salaries, including extended contract salary and Career Ladder supplement.

---

[20] The tests also include Science in grades 3, 7, and 10; and Social Studies in grades 4, 8, and 11. By 2006, to comply with the No Child Left Behind (NCLB) Act of 2001, Missouri plans to add Mathematics exams in grades 3, 5, 6, and 7, and Communication Arts (Reading) exams in grades 4, 5, 6, and 8.

[21] DESE uses the following categorizations in collecting information about Career Ladder responsibilities: parent contact, student tutoring, other student contact, curriculum development, professional development, other instructional improvement, and all other activities (by district).

c. Summary Comments

A potential weakness of this study would be the reliance on inter-district comparisons. Differences in district characteristics introduce variation that would reduce the precision of impact estimates. District-level characteristics, which may be confounded with student achievement, include the quality of professional development and other programs offered by a district. Nevertheless, obtaining data on district-level characteristics would improve the ability of this study to estimate more accurately the pay plan's affect. Also, the large number of districts in the state and variation over time would aid identification of program effects considerably.

The study could be relatively easy and straightforward, and could be completed quickly due to the availability of data and the apparent cooperativeness of the state contacts. In addition, Michael Podgursky of the University of Missouri is a consultant to our team. Dr. Podgursky can provide additional information and contacts to facilitate obtaining and understanding data for the study. A study of this program could shed some light on the links between student performance and teacher incentives for extra work and professional development based on a program that has proven its long-term viability.

6. ARKANSAS HIGH PRIORITY DISTRICT BONUS PROGRAM

a. Overview of Program

Goal. The goal of this program is to help small, rural districts attract and retain qualified teachers. The program is statewide, but it is targeted to districts with enrollments under 1,000 in which more than 80 percent of the students are from low-income families — districts found mostly in the Mississippi Delta region.

Program Structure and Operation. Teachers in Arkansas receive bonuses for working in a "high priority" or "high need" district. High need is defined as having 80 percent or more of the student body eligible for free or reduced price lunch and having fewer than 1,000 students enrolled (based on average daily membership in the previous school year). The program was implemented as a pilot in eleven districts.

The two distinctive features of this program for our purposes are:

1. ***Focus on Rural Areas***. The eligibility cutoff for districts is based on district size (student enrollment), which in Arkansas means that it also targets rural areas.
2. ***"Pure" Recruitment and Retention Model***. Bonuses are awarded simply for accepting or returning to a teaching position in the district, regardless of teacher characteristics or performance. In that sense it is a "pure" recruitment and retention program, and does not include performance-based pay or pay for skills or extra work.

Eligibility. Eligibility for receiving bonuses is based on school district characteristics, specifically size and poverty. As mentioned above, districts must have:

- At least 80 percent of students eligible for free or reduced price lunch
- Fewer than 1,000 students enrolled in the 2003-2004, year before the program was first implemented. At the same time the bill was passed in the legislature, another bill required districts with less than 350 students to consolidate or close, so the effective eligibility is for districts between 350 and 1,000 students.

Districts that meet both of these criteria are placed on a list of high priority districts and included in the recruitment and retention bonus program. While we identified 7 districts that appeared to be eligible based on data provided by the Arkansas Department of Education (ADE), the state officials reported that 11 districts were included in the program. Six of those 11 were on the list of the eligible districts that we identified. An additional three were just below the lower threshold for district size, with between 300 and 350 students, and the remaining two were below the poverty threshold, with 79 and 66 percent of their students eligible for free or reduced-price lunch. That leaves one of the eligible districts that was not invited to participate in the state program. While we called many of the districts to verify their participation in the program, we are still seeking clarification from ADE officials of reasons why five near-eligible districts and

one clearly non-eligible district were included in the program and why one seemingly eligible program was excluded.

Nearly all certified teachers in the pilot districts are eligible for bonuses. The state agency regulations list as eligible those "certified personnel who spend at least 70 percent of the time working directly with students in a classroom setting teaching all grade-level or subject matter appropriate classes, including guidance counselors and librarians."

Award. The size and timing of bonuses depends on whether the teacher is new to the district or is a returning teacher. Teachers who are new to the district receive a signing bonus of $4,000 and a retention bonus of $3,000 per year for each of the following two years, for a total of $10,000 over three years. Teachers already in the district when the program was created receive a retention bonus of $2,000 per year for up to three years. Teachers who leave voluntarily without medical excuse during the year, or before the three-year retention period is over, must return a prorated portion of the bonuses.

Operational Dates. The program began in the fall of 2004 and was expected to be piloted for at least three years, so the first three-year cycle for retention bonuses should end in the spring of 2007. This timing suggests that an early evaluation of the program would be able to estimate its effects only on teacher retention or recruitment outcomes after one or two years.

The long-term viability of the program depends on funding from the state legislature. The funding that authorizes the program still must be renewed for the program to continue. ADE administrators have told us that the program's chief supporter in the legislature recently left office because of term limits, but advocates believe that favorable early outcomes will help their efforts to renew and continue the program.

Policy Relevance. Recruiting and retaining teachers is a common challenge for school districts in rural areas because of geographic isolation and poor local economies. Many districts and states around the country use monetary incentives to attract and retain teachers in high-need geographic areas. For example, there is a federal rural educator program with funds that often are used by grantee states to provide such incentives. Other well known programs exist or are being proposed, such as several in other parts of the Mississippi Delta and Virginia. Arkansas provides a relatively clean example of a program with well-defined criteria for participating districts. Many similar programs in other states combine teacher recruitment incentives with training, alternative certification, and inservice professional development, which makes it difficult to determine how much of a program's impact is due to the bonuses. Furthermore, the Arkansas bonus amounts are substantial, totaling $10,000 and $6,000, respectively for new and returning teachers.

Prior and Ongoing Research. No evaluation of this program has been done, although the original legislation requires a "comprehensive evaluation" by September 30, 2006. Our best understanding is that ADE staff are preparing some tabulations to satisfy this requirement, but we have not received direct confirmation of their methods for presenting the information.

b. Overview of Possible Study Design

Research Questions. The research would address the following questions:

- Does the presence of a recruiting bonus raise the quality of new teachers in high-need areas?
- Does the presence of a retention bonus raise the retention rate of high-quality teachers in high-need areas?

Identification of Program Effects. The Arkansas program is particularly well suited for a regression discontinuity design, where the identifying index variable has two dimensions: district size and poverty rate (percentage eligible for free or reduced price lunch) and one of those (district size) has two thresholds. We also will exploit the variation over time by including data from the pre-implementation period on both program districts and comparison districts. Table 5 shows the distribution of Arkansas school districts along both dimensions of eligibility criteria.[22] The table suggests that a regression discontinuity design is promising because there are school districts that lie both above and below, near and far from the eligibility thresholds. This includes districts that meet one eligibility criterion but not the other.

Not shown in the table is the information from subsequent years. District eligibility was based on just the year prior to implementation (2003-2004), but some districts that were ineligible in that year had the same characteristics in subsequent and/or previous years, suggesting that their eligibility status was random, an artifact of year-to-year fluctuation in the student population. This use of a single base year for eligibility determination makes the regression discontinuity argument especially compelling as a type of natural experiment.

One caveat to bear in mind is the fact that the discontinuity may be "fuzzy" rather than sharp. In other words, we identified districts that fall on the ineligible side of the cutoffs but that were included in the program as pilot sites. We will explore the degree to which this could reduce precision of or introduce bias into the estimation of program impacts. One possibility is that the true lower bound cutoff for enrollment was 300 students and not 350.

Variables and Measurement. Measuring outcomes for programs with a recruitment focus is challenging. Ideally, we would like to know the number of teachers hired, divided by some measure of the demand or need for teachers (by category), as well as the district's average recruiting cost per vacancy. For recruitment, the relevant variables we could obtain are the percentage of classrooms taught by highly qualified teachers (reported by the state every year by October 15) and the rate of out-of-field teaching, which can be measured using counts of waivers

[22] Recent information from ADE suggests that many districts in Arkansas that met the eligibility criteria were arbitrarily excluded from the program in its first two years so that the state could focus on a select set of pilot districts. We are working with ADE to determine whether districts that met the eligibility criteria but were not included in the pilot would be appropriate for comparison group in addition to the districts we will use in the regression discontinuity deign.

TABLE 5

NUMBER OF ARKANSAS DISTRICTS BY POVERTY AND SIZE (2003-2004)

	Percent of Students Eligible for Free/Reduced-Price Lunch						
Student Enrollment	<40	40 to 49	50 to 59	60 to 69	70 to 79	80 to 89	90 to 100
<350	4	3	11	22	13	5	8
350 to 1,000	11	35	39	30	5	4	3
1,000 to 1,500	6	10	17	4	1	0	0
1,500 to 5,000	13	19	16	7	4	0	5
>5,000	5	5	3	1	1	0	0

Note: Cells in dark shaded area (7 districts) meet the eligibility criteria for "high need." Cells in light shaded area (32 districts) are highlighted to show how many are nearly eligible.

by school and district for out-of-field teaching. For retention, we will seek data from the state that will allow us to calculate the continuation (retention) rate for teachers by years of experience, grade level, subject, and district.

Because Arkansas has not switched to annual testing in every grade until very recently, we might not be able to examine the effects of the program on achievement. However, student test score data are available if we were to consider using different subject tests as pretests, e.g., Grade 3 reading as a pretest measure for Grade 4 mathematics. Such an approach makes sense if performance across subject areas is highly correlated.

Data Availability. ADE maintains three databases that can provide the student and teacher information needed for this study:[23]

- Arkansas Statewide Information System (data on instructional expenditures, staff characteristics, and student characteristics)
- Student Performance Database (data on student test scores)
- Arkansas Professional Licensure System (data on teacher licensure)

The databases are housed at the University of Arkansas' National Office for Research, Measurement, and Evaluation Systems (NORMES). School-level aggregate information can be obtained on the Internet. Requests for individual level data would have to go through the ADE, and the researchers would interact with NORMES staff. Our conversations with ADE staff and

[23] Detailed information on these data come from the Southwest Educational Development Laboratory (2004).

researchers at the University of Arkansas (Gary Ritter and Jay Greene) suggest that obtaining data in this way is feasible.

c. Summary Comments

While analysis of retention outcomes is straightforward if the hiring data are available, studying any type of teacher recruitment bonus program is challenging because the direct impacts are difficult to measure. Such programs aim to change the behavior of a pool of individuals—would-be teachers—that is difficult to identify *a priori*. However, it is possible to look at the eligible schools or districts and determine whether their staffing, and possibly student achievement outcomes, are better as a result of the program's existence. This depends on having good measurement of teacher hiring, especially of teacher quality or qualifications.

Arkansas provides perhaps the best test case available for this type of program. The size of the bonuses is substantial. The criteria for eligibility and award of bonuses are clear. There are arbitrary eligibility cutoffs for districts, below and above which we can compare outcomes. The data in Arkansas appear accessible and comprehensive enough to construct meaningful outcomes and conduct a set of data analyses that will be informative.

There also are some drawbacks to keep in mind, however. The program is not yet mature. Because the program is still in its start-up phase, ADE may decide to make changes over time, weakening the link between our findings on early implementation with later implementation policies or may fail to renew the program at all. Another drawback is that eleven districts is a small number of pilot sites. Nevertheless, we believe that the program is designed well enough, and has sufficient data on pilot districts in the pre-implementation years and comparison districts in the pre- and post-implementation years, that it merits the proposed investigation.

7. PALM BEACH COUNTY (FLORIDA) TITLE I SIGN-ON INCENTIVE PROGRAM

a. Overview of Program

Goal. The goal of this program is to recruit and retain qualified teachers in Palm Beach County's schools that are eligible for school-wide Title I programs.[24]

Program Structure and Operation. Palm Beach County (PBC), Florida, offers incentives for teachers in schools that are eligible for implementing school-wide Title I programs (hereafter, Title I schools): signing bonuses for new hires, and tuition reimbursement for teachers taking courses. New hires that teach in-field in one or more core subject areas or critical shortage areas receive $5,000 signing bonuses if they commit to teach in a Title I school for four years. There are teacher shortages in almost all subject areas in Title I schools. The "core subject/critical shortage areas" in these schools are: language arts, math, science, social studies, elementary education, special education, and reading.

Also, the district subsidizes coursework (tuition and books) for eligible teachers in Title I schools. The teachers may take courses to earn advanced degrees (Masters, Specialist or Doctorate). Eligible teachers who were assigned to teach core subject/critical shortage areas, and were teaching out-of-field in the 2002-03 or 2003-04 school years are reimbursed for taking courses to comply with their out-of-field agreements if they commit in writing to stay at a Title I school for the duration of their study, plus two years. Similarly, the district awards tuition reimbursement to in-field teachers who take courses to become certified in a core subject/critical shortage area, and commit in writing to teach at a Title I school for the duration of their study, plus two years.[25]

Additionally, the district gives priority to teachers in Title I schools when they apply for staff development workshops.

The two distinctive features of PBC's program for our purposes are:

1. ***Focus on Title I schools***. Eligibility for the intervention is determined exclusively by a schools' status as being eligible for schoolwide Title I assistance. This criterion

[24] Schools nationwide may receive federal Title I funds to operate a targeted assistance program or a school-wide program. A targeted-assistance school must focus its services on children identified as "failing, or most at risk of failing, to meet the state's challenging student academic standards." According to federal guidelines, a school must have a child poverty rate of at least 40 percent to choose to operate a school-wide program. In a school-wide program, most federal, state, and local funds are consolidated to upgrade the entire education program of the school. In such schools, Title I is no longer a distinct program but is integrated into the regular educational program of the school. See http://www.edweek.org/rc/issues/title-i/.

Palm Beach County only has school-wide Title I programs, and schools are eligible for the programs if they have a child poverty rate of at least 50 percent. The county has no targeted-assistance schools.

[25] In addition to focusing on teachers who are not teaching in a core subject/critical shortage area, this incentive is also for teachers who have certification in one core subject/critical shortage area, but may wish to become certified in another.

has an arbitrary threshold of 50 percent disadvantaged students that creates a natural experiment. That is, schools just above and below this threshold should be nearly identical to each other in every way on average except for their eligibility for the bonus.

2. ***"Pure" Recruitment and Retention Program***. The district awards the incentives to teachers in Title I schools, regardless of teacher characteristics or performance. In that sense it is a "pure" recruitment and retention program, and does not include performance-based pay or pay for skills or extra work.

Eligibility. PBC's highest-poverty schools, with greater than 50 percent of students eligible for free or reduced-price lunch, are eligible for implementing school-wide Title I programs, under which the Sign-On program operates. In the 2005-06 year there are 120 Title I schools, out of a total 223 schools in the district, where new teachers can commit to teach, and receive signing bonuses.[26] Table 6 lists the number of schools by poverty status, with shading to indicate cells with eligible schools and near-eligible schools. To receive signing bonuses, the new teachers must hold, or be eligible to receive a Florida teacher certificate. They must have an active contract with the district and be teaching in-field at the time the bonuses are due them. To obtain tuition reimbursements, teachers had to sign up by September 30, 2003.[27]

TABLE 6

NUMBER OF PBC SCHOOLS BY POVERTY LEVEL (2005-2006)

Percent of Students Eligible for Free/Reduced Price Lunch					
Ineligible		Not Quite Eligible	Borderline Eligible	Eligible	
<30	30 to 39	40 to 49	50 to 59	60 to 69	70 to 100
66	23	14	23	20	77

Note: Cells in dark shaded area (120 schools) include schools that are eligible for the sign-on bonuses. Cells in light shaded area (14 districts) are highlighted to show how many schools are nearly eligible.

Award. The district pays the signing bonuses in three installments. Newly hired teachers receive the bonuses as follows: one-third of the bonus amount ($1,666.67) 30 calendar days into

[26] These include all K-12 educational institutions, such as magnet schools and charter schools. For a study we will most likely focus on the regular public schools. Each year, since 2003-04, the number of Title I schools has increased as the poverty levels in these schools have increased above the 50 percent cut-off. For example, there were 13 new Title I schools in 2005-06, as compared to 2004-05. Our initial research indicates that there were 96 Title I schools in 2003-04. No schools have moved out of Title I status. We have requested information about the poverty levels of schools in the district in 2003-04, and will be following up to obtain this information as needed.

[27] The district had a limited budget for the program, and eligible teachers could only obtain reimbursements if they signed up for the program by September 30, 2003.

their first trimester (after they complete a New Employee Orientation); the second third of the bonus within 30 days after the third trimester; and the final third of the bonus at the end of the first trimester of the second year.

If a teacher ends his/her employment at a Title I school before the end of the four-year commitment period, he/she pays back a prorated portion of what has been received from the district. For example, if the teacher leaves after three years, he/she will repay a quarter of the bonus ($1,250) to the district. If a teacher transfers to another Title I school he/she remains in the program, and does not have to pay any money back. A participating teacher may take a leave of absence, but must satisfy the commitment upon returning from leave. Also, a participating teacher who leaves the district may be rehired by the district, but is not eligible for a signing bonus when he/she returns.

Operational Dates. The district has awarded signing bonuses and tuition reimbursements for three years: 2003-04, 2004-05, and 2005-06. PBC and the Classroom Teachers Association of the district are currently modifying the incentive program, to give bonuses only in the areas of special education, reading, science, and math. They have yet to determine the revised amount of the bonus.

Policy Relevance. Recruiting and retaining teachers is a common challenge, especially for high-poverty schools, and many districts and states around the country use monetary incentives to attract and retain teachers to such schools. PBC's program is relevant because it provides a clear example of a program with well-defined criteria for participating high-poverty schools.

Prior and Ongoing Research. No evaluation of this program has been done. However, the district has kept data on how many teachers were eligible to receive the incentives, how many accepted, and how many remained in a Title I school. We have requested this information, and we will be following up to obtain the information.

b. Overview of Possible Study Design

Research Questions. The research would address the following questions:

- Does the presence of a recruiting bonus and tuition reimbursement program raise the recruitment success rate of new teachers in high-poverty schools?
- Does the program increase retention of teachers in these schools?
- Does the program raise the quality of teachers in these schools?

Identification of Program Effects. We plan to use a regression discontinuity design, which uses the cutoff of the school poverty rate (percentage eligible for free or reduced-price lunch) for school-wide Title I program eligibility. The program schools are those that were just above the cutoff, whereas comparison schools are those that were just below the cutoff, and missed being eligible for school-wide Title I as a result. We also propose to study program effects by

exploiting variation over time, and including an analysis of data from the pre-implementation period on both program schools and comparison schools.

Variables and Measurement. The three key outcome variables would be recruitment success rates, retention rates, and levels of teacher quality. We will measure recruitment success rates by using data on teaching vacancies and numbers of teachers hired to fill the vacancies in schools. We will note other related variables, such as the rate of out-of-field teaching. We will measure retention using data on teacher mobility by years of experience (noting time in school/district), grade level, subject, and school. We will use teacher certification (and degree, if possible) as a measure of teacher quality. We will also analyze student test scores as an outcome of interest, related to teacher quality. We will control for relevant student and teacher demographic characteristics, and school characteristics, such as school size and type of school (elementary, middle, high school).

Data Availability. The Florida Department of Education (FDE) has comprehensive student and teacher data in its K-20 Education Data Warehouse (see Florida's data information under the Data Availability section for TAP).

c. Summary Comments

Palm Beach County presents a good opportunity to study recruitment and retention programs. The size of the bonus is substantial. The high-poverty criteria for program eligibility are nationally relevant; there are school-wide Title I eligibility cutoffs for schools, below and above which we can compare outcomes. The data in Florida appear accessible and comprehensive enough to conduct informative data analyses and make meaningful inferences.

However, there are some challenges to consider. The effects of the program could be conflated with any effects of other programs that are implemented in PBC's Title I schools. While there are a few programs that began operating in PBC a year or more after the bonus and reimbursement program began, staff from the PBC federal grants office have indicated that most Title I schools use their school-wide funding exclusively for the signing bonuses and tuition reimbursements. PBC has identified only a few Title I schools that have implemented other programs in addition to the teacher incentives. Additionally, most of the high-poverty schools in the county that are not eligible for school-wide Title I funds have no other programs in place. Another drawback to a study of this program is that it will be restricted to one school district. District-specific effects may reduce the generalizability of study findings. Nevertheless, we believe that the policy relevance of PBC's Sign-On program, and the availability of data in Florida outweigh the drawbacks, and make this program a strong candidate for study.

D. SAMPLE SIZE ADEQUACY

We believe the seven recommended study ideas would each contribute useful new information to the debate on teacher pay reform. In some cases we would have to qualify our inferences and note carefully the degree of confidence we have in the findings. The results of a pilot study with 50 teachers in 8 schools over two years, for example, may depend somewhat on the particular schools that were included in the pilot, but the study would be a critical data point in future attempts to learn from the variety of teacher pay reform efforts, whether by meta-analysis or less formal research synthesis.

We have not yet conducted formal power analysis for this set of study opportunities because we are continuing to gather information that would make such power analysis realistic and informative. However, our ability to detect small program impacts also will vary, because the units of analysis and sample sizes vary. Many of the programs whose feasibility we assessed are small-scale pilot demonstrations, so it will be difficult to make statistical generalizations beyond the schools and districts included in the programs.

Though detailed power calculations are not possible for each study at this time, Table 7 lists various aspects of the expected samples that will be available for analysis. We will analyze each site at the teacher, school, or district level, depending on the units across which the program was implemented.[28] With the exception of the Missouri career ladder program, most programs began operating recently. We can obtain at least four years of data for each site, through either the Internet, data centers such as NWEA or the North Carolina Education Research Data Center, or districts and states. Table 7 shows the number of units implementing the program and the number of years we would be able to observe the program in operation. It is important to note that the intervention being studied is nearly always the eligibility for receiving an award, i.e. being subject to the incentive, not whether an individual teacher or school actually received an award, since it is the eligibility that is hypothesized to affect behavior. Expected program group size ranges from six schools in the case of the Charlotte-Mecklenburg Pay for Performance pilot to thousands of schools in the case of the California award program.[29] In most cases we will be able to exploit variation over time by observing the program group for at least two years; however, the duration of program implementation and the years of data availability limit our observation of the program group to only one year for the Charlotte site and for some of the TAP schools. Finally, Table 7 lists information on the possible comparison groups. The analysis of each program typically will involve comparing schools or districts participating in the program to similar nonparticipating schools or districts. The size of these comparison groups ranges from dozens to potentially thousands of similar units. An exception occurs for Cincinnati; as we propose identifying the effects of the Cincinnati pay system by observing a panel of teachers over time, there is no explicit comparison group design for this program.

[28] To extract the most information from the data and yield correct standard error estimates, we expect to estimate multilevel models in most cases, with the levels being years, students, classrooms, schools, and districts. In most cases, it will be feasible and appropriate to model two or three levels.

[29] For most of the sites, data may be available at lower levels of aggregation, such as classrooms within schools or schools within districts.

TABLE 7

EXPECTED CHARACTERISTICS OF SAMPLES AVAILABLE FOR ANALYSIS

Site	Main Unit of Analysis	Year Program Began	Number of Years of Data Available	Size of Program Group	Number of Years Observe Program	Possible Comparison Groups	Size of Comparison Group
Category 1 (Pay for Performance)							
California	School	1999	6	1,300 to 4,000 schools	2	Similar low-performing schools	1,000 to 5,000 schools
Charlotte-Mecklenburg, NC	School	2004	10	6 schools*	1	Matched comparison schools All other FOCUS schools in district	42 schools
TAP: Arizona	School	2000	6	7 schools[a]	5	Similar non-TAP schools	Max of 407 schools
TAP: Arkansas	School	2002	6	14 schools[a]	2	Similar non-TAP schools	Max of 272 schools
TAP: Colorado	School	2002	7	15 schools[a]	3	Similar non-TAP schools	Max of 934 schools
TAP: Florida	School	2001	7	22 schools[a]	4	Similar non-TAP schools	Max of 2,666 schools
TAP: Louisiana	School	2003	6	6 schools[a]	2	Similar non-TAP schools	Max of 684 schools
TAP: Minneapolis, MN	School	2004	7	3 schools[a]	1	Similar non-TAP schools	Max of 43 schools
TAP: S. Carolina	School	2001	7	11 schools[a]	4	Similar non-TAP schools	Max of 599 schools
Category 2 (Pay for Knowledge, Skills, or Extra Work)							
Cincinnati, OH	Teacher	2002	4	162 teachers	4	(None; identification over time from panel of teachers)	(None)
Missouri	District	1986	6	309 to 338 districts	6	Matched districts among nonparticipating districts	Max of 186 to 215 districts
Category 3 (Pay for Filling a Need)							
Arkansas	District	2004	4	11 districts[a]	2	Similarly sized, high-poverty districts	20 to 50 districts
Palm Beach County, FL	School	2003	7	96 to 120 schools*	4	Similar high-poverty schools in the district	Max of 103 schools

[a]Data may be available at lower levels of aggregation, such as classrooms within schools or schools within districts.

E. POTENTIAL COSTS OF CONDUCTING STUDIES

To aid IES in deciding which programs to study using secondary data, we estimated the potential costs associated with different proposed analyses. Rather than develop separate and specific cost estimates for studying each of the seven programs, we developed three general cost models based on important differences in the way we would have to obtain the data. We felt that these differences would be a major factor in driving overall costs, and that the differences between these categories are likely to be larger than the differences in costs for specific programs within each category.

The programs fell into three groups that correspond to low-, medium-, and high-cost options (see Table 8). For four of the programs (in Arkansas, California, Missouri, and one of the state analyses for TAP) we believe that we will be able to obtain most of the necessary data quickly and easily, simply by downloading it from one or more state websites. For three of the states that may be analyzed as part of a study of TAP, Arizona, Louisiana, and South Carolina, some of the data also are available on the Internet, but rather than downloading an existing database, we would have to click through many separate pages and extract data one school or one district at a time (or simulate that effort with an automated program). Finally, for the remaining four programs (in Palm Beach County, Florida; Cincinnati, Ohio; Charlotte-Mecklenburg, North Carolina; and several of the states we identified for a study of TAP) we would have to submit requests to state agencies or data warehouses and wait for them to extract, compile, and send us the data we need.

TABLE 8

ESTIMATED COSTS OF CONDUCTING SECONDARY DATA ANALYSES

Data Source and Data Collection Model	Program Locations[a]	Estimated Costs for Illustrative Example
Internet Download	Arkansas California Missouri TAP (AR)	$175,000
Internet Click-Through	TAP (AZ, LA, SC)	$185,000
State Agency or Data Warehouse	Florida Ohio North Carolina TAP (CO, FL, MN)[b]	$225,000

[a] Some programs are identified here by location for brevity only. See Table 1 for full names of the programs.
[b] The costs of conducting the TAP study would be up to three times the amount shown in the illustrative examples, depending on how many states were included for analysis and whether any of the other states already were being studied.

We generated the cost estimates by budgeting the effort based on a set of assumptions about research staff hours and other resources required to revise and submit a formal analysis plan, obtain the data and clean it, process it for impact analysis, conduct the impact analysis, and write a report. These tasks are described in more detail below.

It is important to recognize that the cost estimates shown in Table 8 are illustrative only, and make many assumptions that will have to be updated in light of the following: programs and jurisdictions actually selected for further study, any subsequent clarifying information we receive from state and local education officials, researchers who have worked with the data, and any revisions we make to the analysis plan.

Task 1: Analysis Plan. This task would entail formulating the appropriate statistical models and developing the identification strategy, drafting a formal analysis plan, developing sample figures and table shells, and revising the plan in response to feedback from IES and consultants. We assume that the costs for developing an analysis plan would be fairly similar, no matter which kind of program is being studied.

Task 2: Obtain Data. This task involves steps such as identifying the districts and schools for which data are needed; submitting requests for the data and gaining permission to access and use it; and, in some cases, downloading it from a website. It includes drafting and submitting research proposals to state or local education agencies, if necessary, obtaining Institutional Review Board approval, and paying fees to data centers as required. For example, we may need the North Carolina Education Research Data Center housed at Duke University to prepare confidential data to our specifications, including matches of student information to teacher background data, all of which would require clear communication and permission to proceed. The Data Center has a typical data-handling fee of $850 per day. In addition, we obtained illustrative cost estimates from the NWEA, which maintains an extensive database of student test score data—usually from its computer adaptive tests—for thousands of schools around the country. Their estimates indicated that the processing fees for a typical data extraction would be approximately $1,200 to $4,400, plus a variable amount based on the number of students whose records were being requested. In one example, the variable component of the fee can be $5,000, for accessing the records of 2,650 students, or as high as $50,000 for obtaining larger samples.

Task 3: Process Data. This task would entail entering data into a useable format such as SAS, aggregating data from the school to district level as necessary, merging multiple databases obtained from different sources if necessary, creating comparison groups, constructing variables, clarifying our understanding of the data through communication with the data providers, and resolving data problems and anomalies. Depending on whether merges can be done using consistent and unique ID codes, or whether we will need to conduct name merges, clean up mismatches, and check for errors in ID coding, this task could be either simple or complex.

Task 4: Analysis and Reporting. This task involves analyzing quantitative data, planning and conducting a focused implementation analysis, writing up the results, briefing IES, and revising the draft. We assume that the costs for analysis and reporting will be similar for each of the programs being studied. This task includes travel expenses for a qualitative researcher to visit the district or state to conduct interviews related to program implementation or help obtain data.

These estimates are preliminary and can serve as guidelines for planning, but should be updated with information specific to each analysis before committing resources.

F. SUMMARY AND RECOMMENDATIONS

Having conducted an initial feasibility study, we believe that secondary analyses using existing data to estimate the effects of teacher pay reforms on teacher performance, recruitment, and retention is not only feasible but also likely to yield important insight into the effectiveness of teacher incentive policies. We have identified several instances where a natural experiment can be said to have occurred, where an arbitrary decision rule for program eligibility creates groups of ineligible schools or districts that look as if they had been randomly assigned to a control group. Below we summarize our efforts and offer recommendations on how to select candidates for further study, although at this time we do not have a specific recommendation on which programs should and should not be selected.

The feasibility study began with a national search for teacher incentive programs that would be good candidates for further study and conducted interviews and document reviews to assess their feasibility. We grouped the programs into three categories based on how the programs rewarded teachers, whether for performance; knowledge, skills, or extra work; or for teaching in hard-to-fill areas. In consultation with IES, we then identified the most promising and feasible candidates within each category. Finally, we conducted more in-depth background research and developed detailed profiles for each of the seven candidates we selected. It is important to note that the seven programs we profiled were deemed the most promising on the basis of information available at the point at which we needed to narrow the field.

Our preliminary cost analysis suggests that approximately three of the study opportunities would be feasible to study within IES's budget constraint, perhaps two if one of the programs is TAP. Instead of recommending any three in particular or providing a rank ordering, we offer several criteria to help further narrow the decision about which programs would be most fruitful to study:

- ***Policy Relevance***. Has the program been implemented properly? Is the program replicable? Is there demand for information about programs such as this? Would a new study be necessary to assess the program's effectiveness, or has the program already been sufficiently studied?
- ***Feasibility***. Are there data that can be used to study the program and its key outcomes? Are the data comprehensive and usable? Is the cost of obtaining the data reasonable? Can the data be obtained and analyzed in a timely manner?
- ***Design***. Is there a viable comparison group or quasi-experimental strategy for estimating program impacts? How convincing is the design? Is the sample large enough to yield meaningful results?

The information gathered to date suggests that all of the programs we profiled in detail could be considered policy relevant. We believe they have all been implemented to a degree that would make them interesting examples to study.

None of the programs we examined has been studied rigorously by independent researchers but nearly all appear to be replicable. Only TAP has been the subject of impact evaluations, and

these were conducted by the TAP Foundation itself. Furthermore, TAP has proven to be a replicable reform model. All the other programs except the California program appear to be within the mainstream of policy proposals that have been put forth in a variety of settings around the country. The California program included extremely large bonuses—as much as $25,000 per teacher—which may reduce the policy relevance of any study of the program. On the other hand, the findings from such a study could act as a useful perspective on all programs by providing an estimate of the effect of bonuses that reach the upper end of the scale.

Policy relevance is an important criterion for deciding the mix of programs to study. One option is to select a single program to study from each of the categories we identified in Section B. Another option would be to focus research resources on just one of those areas, such as pay for performance, where we found the greatest number of programs. One consideration in deciding what mix of programs to study is the Teacher Incentive Fund, a program established by Congress to provide federal grants to states and districts implementing teacher pay reforms. If fully funded at the Administration's proposed level, the program would provide $450 million to states to reward effective teachers and to attract highly qualified teachers to teach in high-poverty schools. Another $50 million would be available for competitive grants to states, districts, and nonprofits for designing and implementing performance based pay. IES may wish to pursue the research opportunities among those we identified that most closely resemble programs that could be funded under this federal initiative.

In terms of feasibility and cost, we believe that extant data can be used to study some combination of two or three of the programs we profiled and that the analysis can be completed within 18 months. We expect that the longest delay in obtaining data would be associated with the Charlotte-Mecklenburg program, for which we estimate a lag of almost one year from the time of student testing to the date when the data are available to the researcher. We should be able to obtain data from other sites more quickly. The cost estimates, which are illustrative only, do suggest that we could study about three programs within the budget constraint of the contract option, consistent with our initial planning assumption.

The feasibility study involved more than identifying programs as candidates for further study. We also began outlining the research design that we would use to carry out each study. The ultimate test of the value of a proposed secondary data analysis is that it convinces stakeholders, policymakers, and researchers that the findings are a meaningful indicator of program effects. Achieving this goal means that there must be a way to estimate the outcomes of program participants relative to what they would have been in the absence of the program, or in other words, relative to the counterfactual.

We have identified a variety of approaches to identifying the counterfactual. Most involve comparison groups of teachers, schools, or school districts and some additional information that can be used to isolate the program's effects from other differences that might exist between program participants and comparison group members. The most prominent strategy is known as regression discontinuity, which treats arbitrary cutoff points in the determination of program eligibility as a natural experiment. Another common strategy is to improve the comparison group design by using covariates (background characteristics of teachers, schools, and districts) and variation over time. In almost all cases, we profiled there is a period before program implementation that can provide a useful benchmark for the outcomes we observe after implementation, setting the stage for a pre-post design. While the results of a pre-post design

alone are often misleading, combining the variation over time with cross-sectional variation (across schools, for example) can provide powerful and convincing evidence of program impacts.

Regardless of which programs are selected for further study, our preliminary conclusion from the feasibility analysis is that every study opportunity comes with challenges, but they each hold promise to make a strong contribution to policymakers' understanding of the effects of teacher incentive programs on important educational outcomes.

REFERENCES

California Budget Project. "Certificated Staff Performance Incentive Awards Miss Their Target." Sacramento, CA, April 2001.

California Budget Project. "What Do the 2000 API Results Tell Us About California's Schools?" Sacramento, CA, March 2001.

Cornett, Lynn, and Gale Gaines. "Quality Teachers: Can Incentive Policies Make a Difference?" Atlanta, GA: Southern Regional Education Board, 2002.

Cornett, Lynn, and Gale Gaines. "Linking Performance to Rewards for Teachers, Principals, and Schools: The 1990 SREB Career Ladder Clearinghouse Report." Atlanta, GA: Southern Regional Education Board, January 1991.

Danielson, Charlotte. "Enhancing Professional Practice: A Framework for Teaching." Alexandria, VA: Association for Supervision and Curriculum Development, 1996.

Driscoll, Donna, and Dennis Halcoussis. "Gains in Standardized Test Scores: An Empirical Investigation." Draft paper. Northridge, CA, January 2005.

Glazerman, Steven. "Teacher Compensation Reform: Promising Strategies and Feasible Methods to Rigorously Study Them." Washington, DC: Mathematica Policy Research, January 2004.

Hassel, B.C. "Better Pay for Better Teaching: Making Teacher Compensation Pay Off in the Age of Accountability." Washington, DC: Progressive Policy Institute, May 2002.

Hough, David. "Teacher Supply and Demand in Missouri: 1999-2000." Springfield, MO: Southwest Missouri State University, 2000.

Johnson, Alvin, Paula Potter, James Pughsley, Calvin Wallace, Eileen Kellor, and Allan Odden. "A Case Study of the Charlotte-Mecklenburg Public Schools School-Based Performance Award Program." Madison, WI: University of Wisconsin-Madison, Wisconsin Center for Education Research, 1999.

Kellor, Eileen, and Allan Odden. "Cincinnati: A Case Study of the Design of a School-Based Performance Award Program." Consortium for Policy Research in Education Working Paper. Madison, WI: University of Wisconsin-Madison, 1999. Available at [www.wcer.wisc.edu/cpre].

Milanowski, Anthony. "The Varieties of Knowledge and Skill-Based Pay Design: A Comparison of Seven New Pay Systems for K-12 Teachers." *Education Policy Analysis Archives*, vol. 11, no. 4, January 2003.

Milanowski, Anthony. "The Criterion-Related Validity of the Performance Assessment System in Cincinnati." CPRE-UW Working Paper Series TC-03-05. Madison, WI: Consortium for Policy Research in Education, University of Wisconsin-Madison, 2003.

Milanowski, Anthony, and Steven M. Kimball. "The Framework-Based Teacher Performance Assessment Systems in Cincinnati and Washoe." CPRE-UW Working Paper Series TC-03-

07. Madison, WI: Consortium for Policy Research in Education, University of Wisconsin, 2003, p. 15.

Missouri Department of Elementary and Secondary Education (DESE). "Recruitment and Retention of Teachers in Missouri Public Schools." Jefferson City, MO, December 2001.

National Commission on Excellence in Education (NCEE). "A Nation At Risk: The Imperative for Educational Reform." April 1983.

O'Day, Jennifer, and Catherine Bitter. "Evaluation Study of the Immediated Intervention/Underperforming Schools Program and the High Achieving/Improving Schools Program of the Public Schools Accountability Act of 1999." Palo Alto, CA: American Institutes for Research, June, 2003.

Podgursky, Michael, Ryan Monroe, and Donald Watson. "Teacher Mobility, Pay, and Academic Quality." Paper presented at the Society of Labor Economists Annual Meeting, May 2002.

Reichardt, Robert, and Rebecca Van Buhler. "Recruiting and Retaining Teachers with Alternative Pay." Aurora, CO: Mid-continent Research for Education and Learning, February 2003.

Rothstein, Richard. "Lessons: Novel Way on Teacher Pay." *The New York Times on the Web*, April 18, 2001.

Sacchetti, Maria. "Awards Ignore Key Factors." *The Orange County Register*, August 13, 2002.

Schacter, John, Tamara Schiff, Yeow Meng Thum, Cheryl Fagnano, Micheline Bendotti, Lew Solmon, Kimberly Firetag, and Lowell Milken. "The Impact of the Teacher Advancement Program on Student Achievement, Teacher Attitudes, and Job Satisfaction." Santa Monica, CA: Milken Family Foundation, November 2002.

Schacter, John, Yeow Meng Thum, Daren Reifsneider, and Tamara Schiff. "The Teacher Advancement Program Report Two: Year Three Results from Arizona and Year One Results from South Carolina TAP Schools." Santa Monica, CA: Milken Family Foundation, March 2004.

Snipes, Jason, Fred Doolittle, and Corinne Herlihy. "Foundations for Success: Case Studies of How Urban School Systems Improve Student Achievement." New York: MDRC, 2002.

Solmon, L.C., K.F. Firetag, and T.W. Schiff. "Improving Student Achievement: Reforms That Work." A volume in the Milken Family Foundation Series in Education Policy. Santa Monica, CA: Milken Family Foundation, 2005.

Southwest Educational Development Laboratory (SEDL). "Investigation of Education Databases in Four States to Support Policy Research on Resource Allocation." Austin, Texas: SEDL, December, 2004.

Tobias, Justin L. "Assessing Assessments of School Performance: The Case of California." *American Statistician,* vol. 58, no. 1, 2004, pp. 55-63.