Teacher Compensation Reform: Promising Strategies and Feasible Methods to Rigorously Study Them

Final Report

January 2004

Steven Glazerman

Submitted to:
U.S. Department of Education
Institute of Education Sciences
555 New Hampshire Avenue, NW
Room 308, Building 2
Washington, DC 20208

Project Officer:
Elizabeth Warner

Submitted by:
Mathematica Policy Research, Inc.
600 Maryland Ave., SW, Suite 550
Washington, DC 20024-2512
Telephone: (202) 484-9220
Facsimile: (202) 863-1763

Project Director:
Steven Glazerman
ACKNOWLEDGMENTS

The author thanks Paul Decker for carefully reading the report and providing insightful comments, Carol Soble for expertly editing the report, and Donna Dorsey for skillfully preparing it for publication.
## CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>INTRODUCTION: ALTERNATIVE PAY FOR TEACHERS</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A. GOAL OF THIS PAPER</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>B. FRAMING THE ISSUES</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>1. Challenges Underlying Teacher Compensation</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>2. Goals of Teacher Compensation Reform</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>3. Important Types and Dimensions of Teacher Compensation Reform</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>C. ORGANIZATION OF THE PAPER</td>
<td>7</td>
</tr>
<tr>
<td>II</td>
<td>ALTERNATIVE PAY STRATEGIES IN CURRENT USE</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>A. HISTORY AND LANDSCAPE OF CURRENT PROGRAMS</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>2. Recent Efforts at Teacher Pay Reform</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>B. WHAT DOES THE RESEARCH LITERATURE SAY ABOUT ALTERNATIVE PAY STRATEGIES?</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>1. Implementation Research</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>2. Effectiveness</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>C. LESSONS FOR FUTURE TEACHER PAY REFORM DEMONSTRATIONS</td>
<td>21</td>
</tr>
<tr>
<td>III</td>
<td>SELECTING STRATEGIES TO STUDY IN A RIGOROUS EVALUATION</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>A. DEFINING THE RESEARCH QUESTIONS</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>B. CRITERIA FOR SELECTING STRATEGIES</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>C. MECHANISMS FOR IDENTIFYING STUDY SUBJECTS</td>
<td>29</td>
</tr>
<tr>
<td>IV</td>
<td>DESIGN OPTIONS FOR A RIGOROUS EVALUATION</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>A. COUNTERFACTUAL OF INTEREST</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>B. MEASURING OUTCOMES</td>
<td>34</td>
</tr>
<tr>
<td>Chapter</td>
<td>Page</td>
<td></td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>C. UNITS OF INTERVENTION</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td>D. MODE OF TREATMENT ASSIGNMENT</td>
<td>41</td>
<td></td>
</tr>
<tr>
<td>1. Experimental design (random assignment)</td>
<td>41</td>
<td></td>
</tr>
<tr>
<td>2. Quasi-experimental designs</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>3. Alternatives to the impact evaluation paradigm</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>V. PUTTING IT ALL TOGETHER—ILLUSTRATIVE EXAMPLES OF EVALUATION DESIGNS</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td>A. DESIGN A. CLASSICAL DEMONSTRATION PROGRAM EVALUATION (RANDOMIZED EXPERIMENT)</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td>B. DESIGN B. PLANNED VARIATION EXPERIMENT</td>
<td>46</td>
<td></td>
</tr>
<tr>
<td>C. DESIGN C: VALIDITY AND COST OF TEACHER PERFORMANCE MEASUREMENT</td>
<td>47</td>
<td></td>
</tr>
<tr>
<td>D. CONCLUSION</td>
<td>48</td>
<td></td>
</tr>
<tr>
<td>REFERENCES</td>
<td>49</td>
<td></td>
</tr>
</tbody>
</table>
I. INTRODUCTION: ALTERNATIVE PAY FOR TEACHERS

How public school teachers are paid in the United States has been a contentious issue for many years. Critics of the traditional system, which pays teachers solely on the basis of years of experience and educational attainment, claim that the system does not adequately reward and promote good teaching. On the other hand, proponents argue that experience and education are important predictors of teacher performance and that the simplicity, transparency, and fairness of the traditional system make it the only logical choice. To achieve the best of both worlds, educators and policymakers have devised several approaches to reforming teacher pay, ranging from minor modifications to ambitious overhauls. But choosing among different policies remains difficult because the scientific literature on their effectiveness remains extremely limited. The U.S. Department of Education (ED) is exploring options for sponsoring new research to fill the need for better information.

A. GOAL OF THIS PAPER

History has shown that successfully implementing a teacher pay reform policy, much less conducting a rigorous study of such a policy, is a challenging task. Many of the most ambitious and interesting pay reform plans have collapsed within a few years under political opposition or fiscal pressure. At the same time, attempts to study teacher pay reform have failed to yield useful results. Therefore, great care must be taken in framing research questions and crafting research designs.

To begin the process of, first, identifying and prioritizing teacher pay reform strategies for study and, second, outlining methods for conducting such a study, the National Center for Education Evaluation (NCEE), within ED’s Institute of Education Sciences, commissioned this
paper. The paper presents ideas that can guide the departments through the initial steps of defining and conducting a new program of research and formulating a study design.

B. FRAMING THE ISSUES

To frame the issues for the rest of the paper, the next section lists some of the challenges that underlie teacher compensation. This leads to a discussion of the goals of teacher pay reform and a typology of teacher pay reform strategies.

1. Challenges Underlying Teacher Compensation

From the perspective of policymakers, the goal of teacher compensation is to maximize overall teacher performance. It can do this either by (1) causing teachers to increase their effort with respect to the outcomes of interest to the policymaker, an increase that we can call a “productivity effect”, or (2) inducing more talented individuals to enter and remain in teaching positions, a change we can call a “composition effect”. Ideally, the compensation policy would recognize and reward good teaching to produce these effects.

A special challenge in teaching, however, is that identifying and objectively quantifying good teaching is particularly difficult. Observing teachers in the classroom does not easily lead to an objective measure of teaching quality. Observing “output” in the form of student achievement is usually not feasible because the “value added” by teachers—the unique contributions a teacher makes to a child’s learning over the course of a year—is not directly observable. We do not normally know how much a child would have learned if he or she had been assigned to a different teacher. The extraordinary efforts required to estimate teachers’ value added or close correlates of value added are referred to as “monitoring costs”, an enduring problem for employers in many professions besides education.

The main response to this challenge has been to compensate teachers based on factors that are easy to observe and easy to quantify objectively. The result is low monitoring costs but weak
incentives. Teacher pay reformers have tried to solve this problem, with varying success, by aligning compensation with a myriad of possible indicators of teacher productivity or talent. Sometimes the task proves too burdensome or costly. Sometimes the alignment is imperfect, resulting in perverse behavior.

A second challenge is that schools and school districts must compete for new teachers. Districts may wish not only to influence the behavior of teachers already on staff but also to attract and retain the most talented people in both the teaching profession and specific teaching assignments and thereby discourage less able teachers from staying. Thus, when considering the effectiveness of differential teacher pay, we must realize that there already exists differential pay between school districts, between public and private sectors, and between the teaching profession and alternative career paths. A pay reform strategy, and also an evaluation design, would have to recognize that the effectiveness of a school district’s compensation policies depends on the broader labor market in which it is located.

A third challenge to teacher compensation is the importance of collegiality. Teachers often must work together to share curriculum materials and pedagogical insights. They may share the same students for different subjects or at different times of the day, or they may team-teach. Similarly, teachers may pool and use common school resources such as paraprofessionals’ and the principal’s time. Individual rewards in this setting can lead to competitive pressures that damage school climate and undermine collegiality and sharing. Group rewards, however, can present their own problems, by requiring members of the group to monitor each other or by distributing rewards in ways that are perceived as unfair.

2. Goals of Teacher Compensation Reform

The concepts just discussed can be unified in a framework of immediate and final outcomes associated with teacher compensation as shown in Figure I.1. The immediate goals can be to
encourage teachers to work harder, to focus on certain subject areas or students, or to work together more collegially. The incentives for behaving in these ways constitute the school climate. Another set of immediate goals is to raise the quality of the teaching force by attracting and retaining the best candidates. Each of these immediate goals contributes to the intermediate goal of raising the overall average performance of the teaching staff. Increasing performance can be defined in terms of raising student achievement but also perhaps in terms of improving other student outcomes (related to attendance, dropout rates, and disciplinary incidents, for example) or parent satisfaction. Whatever the final outcomes, they can be used to drive compensation to align incentives with goals of the district/policymaker/taxpayer. Different pay reform policies address different links in this chain. Some may target the recruitment of teachers of a certain type. Others may have a more direct impact on school climate and how teachers focus their efforts on the job.

These distinctions are useful for planning evaluation research. A comprehensive evaluation would measure a policy’s impact at each stage in the process. A more focused evaluation strategy may be acceptable if we wish to study one link in the chain, such as the effect of policies on recruiting teachers with subject matter expertise in their field or the effect of policies on teachers’ behavior in the classroom.
3. Important Types and Dimensions of Teacher Compensation Reform

There are many ways to reform teacher pay. Different types of plans involve different approaches and sometimes different goals. Despite some recent exceptions (discussed in Chapter 2), most teacher pay reforms tend to be highly customized such that no two look alike. Consequently, it is impossible to construct a short list of well-developed models for adoption by school districts or for evaluation by researchers. Instead, it is essential to identify a set of important dimensions along which pay plans can vary and use the several dimensions as a typology.

To help focus the discussion, we first distinguish among the four aspects of teaching that can be rewarded under a differential pay plan as shown in Table I.1.

Most teacher pay reform policies can be categorized in terms of what is being rewarded: extra duties or time, demonstrated skills or knowledge, filling a need, or measured performance on the job. It is common for alternative compensation plans, such as that used in the Charlotte-Mecklenberg, North Carolina, school system, to offer a mix of these types of rewards, providing different ways for teachers to qualify for bonuses. Historically, many pay reform programs began with the intent to reward superior performance in the classroom but over time faced pressure to shift the emphasis to more pay for more work or more pay for skills and knowledge demonstrated by activities outside the classroom.

Teacher pay plans can vary along still other dimensions:

**Form of the rewards.** Rewards can take the form of one-time bonuses, permanent salary increments, employee benefits such as favorable home loans or education loan repayment, or nonpecuniary benefits such as honorific titles or favorable assignments.

**Size of the rewards relative to the base compensation.** The new system can (1) entirely supplant the uniform salary schedule and account for 100 percent of a teacher’s salary, (2) add token amounts representing a small percentage of a teacher’s salary, or (3) provide varying sizes of rewards between the extremes.
<table>
<thead>
<tr>
<th>What is rewarded?</th>
<th>Example</th>
<th>Some places it has been tried</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Extra work</td>
<td>Career Ladder or Mentor Teacher Program</td>
<td>Utah</td>
</tr>
<tr>
<td>2. Skills or knowledge</td>
<td>National Board Certification bonus</td>
<td>California</td>
</tr>
<tr>
<td>3. Filling a need</td>
<td>Signing bonus for teaching in high-need schools or for hard-to-fill specialties such as chemistry or biology</td>
<td>North Carolina, Mississippi</td>
</tr>
<tr>
<td>4. Measured teacher performance</td>
<td>Merit pay, pay for performance</td>
<td>South Carolina, Colorado</td>
</tr>
</tbody>
</table>

**Trigger mechanism used to make award decisions.** How is the threshold set for qualifying for a reward? What should be the relationship between the skill or performance measure and the size of the reward itself? How should performance be measured? Performance measurement is the central part, and potentially the Achilles heel, of most pay-for-performance plans. An important distinction can be made between performance measures based on and not based on student achievement. Of those based on student achievement, a type of “value-added” measure is often most desirable if the aim is to reward teachers for the contributions they make to student learning, not the factors beyond their control. Of those not based on student achievement, supervisor ratings are most common. Each measurement type has its advantages and disadvantages. Value-added measures are ideal in that they focus on the outcome that administrators most want to improve, but they tend to be technically difficult to implement and hard to explain to stakeholders. Supervisor ratings, on the other hand, can be easier and less costly to use but cannot guarantee objectivity.

**Who qualifies for rewards.** Bonus or incentive plans can reward individual teachers, teams of teachers, or entire schools. In addition, they can differ in scope and include specific subject-area teachers only, all classroom teachers, or all school personnel.

**Visibility of the program.** Some programs rely implicitly on the positive or negative recognition that comes with being identified (or not) as a highly skilled or high-performing teacher. Other programs choose to operate in the background, with teacher evaluations and bonuses remaining a personnel matter between individual teachers and their supervisors.
Most teacher pay plans can be categorized in terms of these critical dimensions. A more careful look at past and current teacher pay reforms will provide a sense of the relative importance of each dimension.

C. ORGANIZATION OF THE PAPER

The next chapter turns to past and current reform strategies and considers what researchers have learned from them. Chapter 3 revisits the typology of teacher pay reform strategies in light of the available empirical evidence to suggest questions for future research and how to identify and select strategies that would be the object of that research. Chapter 4 discusses the methods that could be used to address the chosen research questions by listing a menu of design options and discussing their tradeoffs. To put the pieces together, Chapter 5 presents three illustrative design templates representing alternative approaches to studying teacher pay reform.
II. ALTERNATIVE PAY STRATEGIES IN CURRENT USE

A new study of alternative teacher pay strategies must build on a foundation of understanding about what the “default” teacher pay system looks like, what strategies have been tried before, and what we have learned to date about existing strategies and their effectiveness. Lessons from past experiences and from the research literature can offer clues about where and under what conditions teacher pay reforms can be tested in the future.

A. HISTORY AND LANDSCAPE OF CURRENT PROGRAMS

Since the early part of the 20th century, most public school teachers have been paid according to a uniform salary schedule, a district formula that takes into account only the number of years of teaching and the highest degree or certificate received. Districts may vary in how they set the initial salary, the size of increments, or the relative importance of higher degrees. They may also include additional pay for additional duties, such as coaching a sports team, leading an after-school activity, or preparing curriculum materials. On occasion, school districts have used signing bonuses, scholarships, and loan repayment to lure teachers to schools or subject areas experiencing shortages. Otherwise, differential pay has been rare in the last 50 years.

Periodic calls for reform, however, have brought about several attempts to institute merit pay, career ladders, pay for performance, aggressive shortage-area bonus programs, loan repayment programs, and other reforms that represent substantial departures from the basic fixed salary schedule. The recent history of teacher pay reform can be broken into two periods, a reform wave in the 1980s and a more recent round of reform. These waves tend to follow a “rise and fall” pattern, with many policies being introduced, tinkered with, and ultimately abandoned or watered down. The reasons for this phenomenon are discussed further below.
The policies that have been tried provided several ways for teachers to qualify for rewards, meaning they typically straddle all or nearly all the categories in Table I.1. It is not always easy to identify the operative components, those that truly drive behavior and rewards, but experience shows that the policies designed with teacher input have mostly de-emphasized student test scores, or combined test score-based measures with other indicators.


The first reform period followed a widely influential report issued by the National Commission on Excellence in Education in 1983. The report, entitled A Nation at Risk, touched off a wave of teacher pay reforms across the country. The report was critical of the status of American education and offered several recommendations on a variety of topics, including the following related to teacher pay:

“Salaries for the teaching profession should be increased and should be professionally competitive, market-sensitive, and performance-based. Salary, promotion, tenure, and retention decisions should be tied to an effective evaluation system that includes peer review so that superior teachers can be rewarded, average ones encouraged, and poor ones either improved or terminated.”

“School boards, administrators, and teachers should cooperate to develop career ladders for teachers that distinguish among the beginning instructor, the experienced teacher, and the master teacher.”

“Incentives, such as grants and loans, should be made available to attract outstanding students to the teaching profession, particularly in those areas of critical shortage.”

In the years that followed release of the report, dozens of states introduced legislation to implement teacher pay reforms, and 23 of those states enacted and implemented career ladder programs (Furtwengler 1995), which tied promotion to measured performance and demonstrated skills. For example, Utah responded with a career ladder program in 1984 that specified several ways that teachers could qualify for more pay: extended contract year, job enlargement, performance bonuses, or incentive funding for shortage areas. The details were left to the
discretion of district officials, who were required to submit detailed plans in response to the state guidelines. South Carolina took a more centralized approach, developing three teacher incentive programs and fully pilot-testing two of them, a bonus model and a campus/individual model, which based pay in part on student performance. The Utah and South Carolina plans remained in place for many years, but most of the career ladder policies enacted in that era were short-lived.

Initiated in 1985, the North Carolina Career Development Program operated in 15 pilot sites. Career ladders in North Carolina were based on teacher performance, and teachers who qualified for promotion received job enlargement opportunities. The state later abandoned the effort because the teacher evaluation system proved too complex to be manageable (Cornett and Gaines 2002). Illinois passed a 1986 law to authorize the State Board of Education to study teacher career compensation programs. The state received 30 program applications and funded seven of them, but then scaled down the funding in each subsequent year until 1990, at which time no additional funds were forthcoming (Cornett and Gaines 1991). The Georgia Career Ladder Program was piloted in five districts in the 1988–1989 school year but never received funding beyond that (Cornett and Gaines 1991).

State legislatures initiated and funded most career ladder programs but accorded individual districts discretion in program design and implementation. Arizona followed the Utah model with 15 pilot districts, including the Mesa Public Schools. Mesa officials developed an optional career ladder program that started in 1986. Participating teachers were promoted along rungs in the ladder based on supervisor evaluations in three skill areas: instruction, management, and counseling (Dickson 1990). Student learning factored into these evaluations as well. Career Ladder applicants set their own plan for student achievement goals, which underwent review and
approval by a trained team of teachers, and then took a self-assessment at the end of the year based on those goals.

Of the myriad career ladder programs and demonstrations, more than half were later discontinued and, by 1995, only four states (Tennessee, Utah, Missouri, and Arizona) had policies that were considered viable and that were funded (Furtwengler 1995). Some of the surviving plans, however, were simply state policies that provided funding for local initiatives. In other cases, the plans came to consist mainly of job enlargement (more pay for more work) rather than performance-based pay, or the programs rewarded such a large percentage of teachers as to provide little pay differentiation compared with the traditional salary schedule. Various researchers have documented and explained the "rise and fall" story (Furtwengler 1995; French et al. 1988; Lane 1989). And, in fact, the dim prospects of teacher pay reforms were already the subject of study by researchers (Murnane and Cohen 1986) just as the 1980s wave of policies was gaining momentum. We return to this issue later in the chapter.

2. Recent Efforts at Teacher Pay Reform

The experience of the 1980s has not dampened new calls for reform. Recent years have seen the introduction of several pay plans. The current universe of pay reform programs is remarkably diverse and constantly changing as programs and policies continually come into existence, undergo rule changes, or disappear; accordingly, no census of programs exists. However, the Consortium for Policy Research in Education's Teacher Compensation Project (www.wcer.wisc.edu/cpre/tcomp) has profiled several programs while the National Council on Teacher Quality (www.netq.org) tracks them, and the Southern Regional Education Board surveys them in occasional reports (Cornett and Gaines 1991 and 2002). ED also has provided periodic reports through the ERIC digests (see Goorian 2000).
The round of reforms that took hold in the 1990s initially reflected a growing interest in school-by-school results and the finding from the earlier wave that individual incentives can undermine school climate. Schoolwide gains were increasingly used to reward schools rather than individual teachers. Cornett and Gaines (2002) describe school reward programs in six states as of the 2001–2002 school year: Delaware, Florida, Georgia, Kentucky, Maryland, and North Carolina. The average amount of the rewards in that year ranged from $18,354 per school in Delaware to $107,368 in Georgia. The percentage of schools receiving awards ranged from 5 percent in Georgia and Maryland to 60.5 percent in North Carolina. Most state plans used student test scores plus some additional measures such as graduation/dropout and attendance rates.

Another trend in the new wave of reform was the appearance of more policies that use teacher incentives for recruiting and retention. Such policies were the more likely reforms found in areas serving disadvantaged students, where the strong economy of the 1990s hurt teacher hiring the most. In 2001, the state of New York initiated the Teachers of Tomorrow Program, which began offering loan repayment and tuition reimbursement for teachers willing to serve in high-need schools or schools under "registration review." The state offered newly hired teachers up to $3,400 per year for loan repayment for the first four years as long as they remained in the high-need school and received a satisfactory rating. Incumbent teachers received a similarly sized reimbursement offer for tuition costs for "coursework taken toward earning Master's degree necessary to achieve New York State permanent certification." Provisional teachers in these schools could also take advantage of a $2,000 stipend for taking courses toward obtaining regular certification.

The Mississippi Critical Teacher Shortage Act of 1998 provided a raft of incentives, including scholarships, loan repayment, moving expenses, housing assistance, and bonuses for
teachers willing to work in the Delta (Cornett and Gaines 2002). North Carolina offers middle and high school teachers an $1,800 bonus if they hold a certification in mathematics, science, or special education and agree to teach in low-performing schools or schools with a high percentage of students from low-income families (Cornett and Gaines 2002). Florida and Georgia offered similar bonuses but have dropped them. California recently canceled a bonus program that gave $20,000 to teachers with National Board certification who agreed to teach for at least four years in low-performing schools, and Massachusetts cut a similar bonus program offering cash for teaching in underserved schools. Recent state fiscal crises are probably the main reason for these cuts.

Programs that reward individual teachers based on classroom performance or demonstration of skills have resurfaced as well. Such policies are often authorized and funded by state legislatures but developed and administered at the district level. Merit pay programs that have received a great deal of attention include the Cincinnati Teacher Evaluation and Compensation System, the Douglas County, Colorado, and Denver Pay for Performance programs, and the Teacher Advancement Program, which was developed by the Milken Foundation and implemented in many settings around the country.

The Cincinnati program was unusual in that it entirely replaced the regular compensation system. Beginning in 2002–2003, teachers were to be paid according to their position on a ladder with five career categories ranging from “apprentice” to “accomplished.” Teachers with 22 or more years of experience were allowed to opt out. The rest were subject to an evaluation based on their skills and knowledge, not student test scores, and placed in one of the salary categories. The evaluations were carried out through classroom observations and portfolio analysis. Movement between categories was determined by demonstrated knowledge and skills; movement within categories was determined by seniority. Teachers could qualify for add-ons to
base pay of $1,000 for National Board Certification, $4,600 for a master's degree, $9,375 for a doctoral degree, and $1,250 for a dual certification. In May 2002, the teachers' union soundly voted to reject the plan for reasons that are difficult to discern. Observers say that teachers may have perceived the evaluation system as insufficiently objective or feared that their pay would decrease as a result of the new system (see article in Education World at www.education-world.com/a_issues/issues374b.shtml).

Added to the list of cancelled performance pay programs was that implemented by the School District of Philadelphia. After the teachers’ union endorsed merit pay in its 2000 contract, Philadelphia officials paid 140 teachers $1,000 each for volunteering to participate in a pilot system that used trained peer assessors’ visits to the 140 teachers’ classrooms to measure classroom performance. The decision not to expand the program districtwide was attributed to the unexpectedly high costs and burden of paperwork, according to reports in Education Week.

A program in Douglas County, Colorado, that has been in place since 1994 shares many characteristics of the Cincinnati pay-for-knowledge plan, but it has survived for far longer than most plans. It supplanted the pre-existing salary schedule to reward teachers for obtaining skills and demonstrating knowledge; however, it also operates with group performance incentives and "responsibility pay," which amounts to paying selected teachers for assuming extra duties. In addition, teachers can volunteer to be considered for a $1,000 outstanding teacher bonus based on a portfolio of their work. The outstanding teacher award is criterion-based so that teachers are not competing against each other for the bonuses. In recent years, the school district has added student achievement as a criterion in measuring performance.

Another Colorado program that has succeeded thus far is the Denver Pay for Performance Program, piloted and evaluated by the Community Training and Assistance Center (CTAC). After conclusion of the four-year pilot in the 16 schools that voted to participate, the Denver
school district appears poised to adopt the program for all teachers and entirely replace the existing uniform salary schedule. The plan, which is scheduled for a vote by the teachers' union in spring 2004, includes a mix of all four components of teacher pay plans outlined in Chapter 1: individual and group bonuses tied to student achievement, salary increases tied to acquisition of skills as measured by degree completion and nationally recognized credentials such as National Board certification, salary increases for performance on a series of "professional practices," and bonuses for service in hard-to-staff assignments such as special education or teaching in "designated schools."

Iowa passed a statewide performance-based pay plan. In 2001, the state legislature funded a program to implement comprehensive evaluations of teachers every three years and use the evaluations to place teachers on a four-step career ladder. The program also provided bonuses to schools that met student achievement goals. In 2002, the state scaled back the program's funding because of overall spending pressure and then, as reported by the Des Moines Register in August 2003 and Education Week in September 2003, slated the program for termination.

Arizona, which has the strongest tradition of successful teacher pay reforms, has adopted a flexible approach to performance-based pay. The state issued 15 guiding principles and let all 232 local districts design and implement unique plans. The Arizona Education Association (AEA) analyzed the designs for 62 of those plans and documented their variation (AEA 2002). The association reported that only six districts implemented the 15 principles as their plans became operational, with over half of the plans considered as evolving. Further, more than half of the plans were based at least in part on student test scores.

One model that has taken hold in five Arizona schools is called the Teacher Advancement Program (TAP). It is somewhat unique in that it was developed by the Milken Foundation with the aim of testing and replicating it throughout the country. TAP is touted by its developers as a
comprehensive school reform model, but its main thrust is market-driven teacher compensation. It also includes several career paths for teachers, teacher accountability based on reviews from trained classroom observers, and value-added measures of student achievement. In TAP schools, the principal enjoys flexibility in compensating teachers differently based on their position on the career ladder as well as on their performance and their ability to fill special needs such as teaching in a high-need school or teaching in a critical subject area such as mathematics or science. TAP schools also advertise out of state and assist new teachers with obtaining certification, or alternative certification as applicable, to help boost the quality of the teaching force. Texas, Arkansas (14 schools), South Carolina (6 schools), Louisiana (5 schools), Colorado, and Indiana are trying out the model. Minnesota recently received a federal grant under Title II of the Higher Education Act to pilot three pay structures: a traditional model, TAP, and a home-grown Minnesota Alternative Teacher Compensation model.

Most pay reform policies are developed by and applied to an entire school district, but the TAP example shows that differential pay is also being implemented at the individual-school level. Moreover, private and charter schools are particularly well suited to adopting their own alternative pay programs. The Vaughn Next Century Learning Center is a charter school in Los Angeles that rewards teachers’ knowledge and skills with bonuses of $300 for expertise in special education inclusion, $400 for technology, $1,300 for English language development, and $1,300 for expertise in literacy (Hassel 2002). A peer assessment review committee rates knowledge and skills on a four-point scale. The school also offers bonuses of up to $2,000 based on student attendance, discipline, parental involvement, and teamwork; bonuses of $2,000 if the entire school meets the state academic performance index goal; and bonuses of $3,500 to $4,000 for assuming extra functions such as grade-level chair, mentor, or peer reviewer.
B. WHAT DOES THE RESEARCH LITERATURE SAY ABOUT ALTERNATIVE PAY STRATEGIES?

So many of the teacher pay reforms enacted over the last two decades operated under "pilot" status that it is reasonable to expect the accumulation of a wealth of knowledge. In fact, despite a considerable volume of research completed to date on the teacher pay reforms discussed above, the studies have yet to address the most important questions with sufficient rigor. The impact of any differential pay strategy on student achievement remains largely unknown. A comprehensive literature review is outside the scope of this paper, but a general picture emerges from a casual examination of the evidence. We have learned a great deal about implementation through qualitative research. We have, however, learned little about impacts of policies on student outcomes, although many observers have tried to tackle the question of why some programs survive and others fail.

1. Implementation Research

Regarding the many attempts to mount teacher pay reforms, case studies abound and tell us much about how the plans were developed, what compromises were involved, how the plans changed over time, and often who participated in the programs when they were voluntary.

Some case studies provide cost information. Peterson (1989) examined two Utah school districts where teacher performance was measured by gathering extensive documentation from multiple sources: student survey, parent survey, principal report, teacher tests, professional activity, peer review, systemic observation, and pupil gains. He quantified the cost of time and materials for students, teachers, and other staff and estimated an average cost of $44.70 per teacher (in 1989 dollars). Even though the study is not recent, it provides a template for analyzing monitoring costs as part of a teacher pay reform.

Other studies provide valuable background information on the relationship between skills that are rewarded and outcomes of interest, such as student achievement (Medley and Coker
1987; Milanowski 1999). The National Board for Professional Teaching Standards, whose certificate program is used in many of today’s pay-for-skills programs, has sponsored a large set of studies to examine the relationship between National Board certification and effectiveness in the classroom. The studies fit into the larger psychometric literature on the validity of teacher assessment instruments. Emerging instruments that are likely to receive much attention include the multipart PRAXIS assessment system created by the Educational Testing Service and other integrated teacher assessment systems that are still undergoing development and refinement.

Still other implementation studies shed light on perverse outcomes that are possible when incentives are manipulated. Gramlich and Koshel (1975) conducted an experiment for the U.S. Office of Educational Opportunity in the early 1970s, when private firms were hired to provide reading instruction and were rewarded in accordance with test score gains. Authors found some evidence that teachers gamed the incentive system:

“In at least one of the sites, teachers concentrated their time on children in the middle of the test score distribution, neglecting those at the top who would advance well on their own (test score gains above a threshold were not rewarded), and those at the bottom, whose test scores would not respond to modest additional amounts of teacher time.”

More recent research has probed questions about whether individual incentives poisoned the atmosphere of cooperation in schools, whether group incentives led to free-riding behavior and resentments, and whether uncertainty about evaluation and compensation caused undue teacher stress (Kelley and Protsik 1997).

2. Effectiveness

No randomized experiments have been conducted to determine the impacts of teacher pay reform on student outcomes. The typical evaluation emphasized more easily measured factors such as attitudes of stakeholders summarized from questionnaires (e.g., Oncida 1990; Schiff 2002) or student attendance (Eberts et al. 2002). The number of programs that have been
subjected to even a quasi-experimental analysis of program impacts is noticeably small. Those that have used comparison groups often relied on a self-selected set of teachers, schools, or districts that elected not to participate and specified weak covariates to adjust for differences. Matched comparison designs have been used to study the Mesa Career Ladder (Dickson 1990), the Merit Schools Program in Florida (Lawther et al. 1990), and the TAP implemented in five Arizona schools (Schacter et al. 2001). The studies produced mixed or inconclusive results about the impacts of the respective programs on student achievement.

The one effectiveness question that has been addressed, if not systematically, is the obvious one: Why do some programs survive and most fail? A widely cited 1986 article by Murmane and Cohen delves into the issue by using insights gained from both interviews with stakeholders and case studies of six anonymous districts. Many other commentators have speculated on the reasons. The typical list includes the following explanations:

- Union opposition
- Budget pressure related to economic downturns
- Poor program design: excessive paperwork or lack of credibility in performance measures
- Costly and too primitive technology for monitoring teacher output

The following are some factors often cited as related to the success of merit pay programs:

- Teacher/stakeholder input in the design process
- High base salaries
- Low visibility for the differentiation of pay
- Wide distribution of rewards, such that most teachers receive something
The above summary suggests that the research literature is strong in some areas and weak in others, providing some useful direction as to where future research can make the greatest contribution.

C. LESSONS FOR FUTURE TEACHER PAY REFORM DEMONSTRATIONS

The history of the merit pay and other pay reforms, along with the body of research just summarized, point to some major challenges for future research on the topic. Two decades of trial and error have provided a treasure trove of ideas and anecdotes, but they have also shown that an initial willingness of a state or school district to experiment with teacher pay reforms does not always lead to a sustained commitment to change. New research efforts could be most successful by considering not only the types of reform that are viable beyond the short-term, but the conditions under which pay reforms can take hold and endure.

It is difficult to know which pay plans districts would make a stronger and lasting commitment to adopt as part of a new study. The knowledge- and skills-based pay programs appear more palatable than pay plans that emphasize student achievement, although heavy demands for documentation or expensive assessments may impede passage of any new programs. Job enlargement and extra pay for meeting shortages may be more politically acceptable.

Ultimately, the willingness of a district to accept differential pay will likely depend on several factors: the involvement of the business community in education reform, the willingness of the local teachers' union to compromise, the extent to which teachers are allowed to participate in crafting the program, the tightness in the labor market for new teachers, and the amount of funding available for rewards. The districts most likely to adopt some type of pay reform or to experiment with new pay reforms could also be those in states such as Arizona or Colorado where the state department of education provides both support for alternative teacher
pay and flexibility to design and implement a program that meets local needs. Or there may be individual districts or charter schools where favorable conditions can be found.

Policy makers at the national level have an interest in schools that serve disadvantaged students, but performance based pay has not lasted long in such schools. In locations like New York City, and the Mississippi Delta, state and local policy makers have tended to opt for recruitment incentives over performance or skill/credential acquisition incentives. Careful planning will be required to craft both an evaluation design and a program design that can survive the challenges discussed here.
III. SELECTING STRATEGIES TO STUDY IN A RIGOROUS EVALUATION

Just as teacher pay reform can take many shapes and sizes, a rigorous study of teacher compensation can follow any one of several approaches. It can focus on a small but important aspect of teacher pay reform, such as the mode for measuring teacher performance or the effects of different-sized bonuses on teachers' career decisions. Or it can evaluate the impacts of a comprehensive pay reform package as a whole relative to the default compensation system. The study can rely on existing models, such as those mentioned in Chapter 2, or it can encourage researchers or education officials to devise their own new teacher pay reform strategies specifically for the evaluation.

This chapter proposes and outlines a process for setting up a new study for teacher pay reform. The first step is to draft some research questions that seem most pressing to policy makers. The next step is to consider the types of pay reforms available and apply a set of criteria for narrowing down the list. Finally, the mechanism by which the study sites are selected will help determine the pay reform strategies that are feasible to study.

A. DEFINING THE RESEARCH QUESTIONS

A study of teacher pay reform can address different types of questions. One question is the typical program impact evaluation question, which begins by identifying a comprehensive package of pay reform and asks about its impacts relative to another pay reform package or relative to the default compensation system, that is, the uniform salary schedule. Examples of this type of question include:

- What impact does the Teacher Advancement Program (TAP) have on student achievement? (The TAP is given as an example, but any intervention can be specified.)
- What impact does TAP have on retention of certified and highly rated teacher?
• What is the impact on recruitment? Are teachers with strong resumes more likely to apply for and accept a position at a TAP school than a similar school with traditional compensation?

This overall evaluation question is the type that the NCEE is typically charged with answering. If such a research question is of interest, then the next task is to identify compensation systems for study.

Before proceeding, however, other types of research questions are worth considering. The first type is a set of questions complementary to the basic impact evaluation question and related to process/implementation and cost-effectiveness. For example:

• How was the teacher incentive program implemented?
• If there were multiple ways that teachers could qualify for rewards, which ways were most often used?
• To what degree did the policy result in differential pay?
• What percentage of teachers opted in?
• What percentage of eligible teachers or schools received the rewards and what was the average award?
• If group incentives were offered, did teachers work together and did they monitor each other?
• Did frictions arise in the school over who was recognized for good performance or skills and who received extra pay?

An implementation study, which is usually a qualitative study, can and probably should be conducted in conjunction with an impact evaluation in order to help the audience interpret the impacts. However, given the large number of existing case studies of program operations and surveys of stakeholder perceptions, a standalone implementation study would probably not make a novel contribution to the literature.

A cost or cost-effectiveness component would be especially useful if the research is to focus on pay for skills/knowledge or pay for performance. Such reforms are predicated on the idea
that the costs of performance monitoring or skills measurement are not prohibitive and that an explicit cost study would directly test that claim. So one might pose questions such as:

- What was the average cost per teacher to determine eligibility for rewards? This includes costs, where applicable, of: taking an exam; collecting and analyzing student test score data; or conducting peer and supervisor reviews, documenting performance, and settling grievances.

- What was the net cost of recruitment bonuses compared to the cost of expanded recruiting and outreach without bonuses?

In addition to the costs of measuring teachers' skills or performance, the incremental costs of compensating teachers under skills or performance systems would be of great interest to the policy community. This type of question might take the following form:

- What was the dollar value of extra incentive rewards paid out through the program?

- How large a fraction of base pay did the typical reward constitute?

Compensation costs in a reward system can be difficult to predict when bonuses are entitlements triggered by teacher behavior.¹ It is essential to understand both cost components—monitoring costs and compensation costs—for the simple reason that fiscal pressure is one of the reasons often cited for the failure of differential pay to survive. Linking the costs of different options to their impacts permits the formulation of statements about cost-effectiveness or cost-benefits.

As discussed, pay reform packages pose a research challenge—whatever the focus of research—because they often do not survive for a sufficiently long period to make their investigation meaningful. Another concern is that, to remain viable, reform efforts often must

¹ Some policies fix the bonus pool, but then the size of the bonuses and/or the number of participants receiving them is an unknown.
undergo change; therefore, an evaluation of a specific program at a specific juncture runs the risk of becoming obsolete before it is even completed. Furthermore, teacher pay reform is a "macro" policy, affecting entire markets rather than individual participants. Thus, a comprehensive program evaluation poses some special challenges that may result in high costs or high risks to the design.

To avoid such risk and to guard against studying complex programs as black boxes, ED could pose more generic "basic research" questions to learn about the mechanisms by which teacher pay reform is supposed to work or to learn about specific aspects of teacher pay reform. For example:

- How responsive are teachers to different sizes and forms of incentives?
- How effectively does the performance measure or skill assessment capture the teacher's "value added" to student achievement?

Such questions can be addressed through experimentation carried out on a small scale. For instance, a planned variation experiment could isolate one or two significant aspects of teacher pay and allow researchers to study their impacts in depth for a sample of individual teachers (as opposed to schools or school districts as the unit of intervention). Or teacher performance measures could be subjected to validation testing whereby an impractical but highly accurate measure serves as a benchmark to assess a set of more practical teacher performance or skill indicators. Chapters 4 and 5 discuss these options at greater length.

B. CRITERIA FOR SELECTING STRATEGIES

Many of the research questions just described are too general unless we insert the name of a particular teach pay reform program or strategy. Identifying a program or strategy to focus on can be challenging, but a reasonable place to start is the list of four teacher pay reform types arrayed in Table 1.1—job enlargement, pay for skills and knowledge, incentives for filling a
special need, performance pay—or a combination. The following three criteria are useful for narrowing the field:

- **Relevance.** For the study to make a contribution to knowledge, the strategy under investigation must be a true alternative to (or departure from) the fixed salary schedule. To be relevant to educators, the strategy must be one that can be implemented in a variety of settings. Finally, relevance has a subjective quality that depends on the needs and interests of education policymakers and stakeholders.

- **Technical Feasibility.** As noted, many teacher pay plans collapse under political or fiscal pressures before they can be fully implemented. Therefore, a demonstration program or intervention that is featured in a federally sponsored study would have to be carefully reviewed for its ability to comply with relevant laws, its likelihood of adoption by educators firmly committed to seeing it through, and its probability of remaining fiscally viable over the duration of the study. The requirements of technical feasibility would mean that diverse stakeholders would need to guarantee their funding and legal commitments. In addition to the feasibility of implementing the program, we must consider the feasibility of conducting a rigorous study. Some teacher pay strategies will be more likely to meet the technical feasibility criterion than others.

- **Research Base.** Before undertaking an ambitious and rigorous impact evaluation, it would be useful to identify some evidence of the strategy’s viability or some evidence that it can lead teachers to change their behavior. In other words, candidates for study should be subjected to the following questions: What do we know? What do we need to know?

The relevance criterion would lead different people to select different strategies, but many would agree that performance-based pay is the most controversial strategy, followed by rewarding skills and knowledge. Less subject to debate are the policies that provide more money for more work or bonuses to meet special needs. For any of these reform types, the extent to which rewards are unequally distributed would arouse more interest. ED should be wary of studying a program in which a very large (or very small) percentage of teachers receives extra
pay; the lack of differentiation would render the program less interesting to inquiries of pay reform.²

As discussed above, the technical feasibility criterion works against attempts to evaluate policies that combine several approaches to teacher pay reform and ask about the various approaches’ impacts on overall performance. The criterion would lead to the rejection of proposed studies that seek to supplant rather than supplement the teacher compensation system. In fact, the more generous the bonus system, the more likely it would elicit the willing participation of teachers, schools, and districts. Unfortunately, more generous also means more costly if the sponsors of the research pay for the intervention. According to the feasibility criterion, the policies that best lend themselves to study are those that provide bonuses for teaching in high-need schools or high-need subject areas. Often the goal of such programs is to attract teachers with minimum (measurable) qualifications such as a teaching certificate and a degree or certificate in the subject area. If so, the study could focus on hiring and retention and avoid the messy and costly task of measuring the value added to student achievement.

The third criterion works in a different way. In consideration of factors such as what is known and what needs to be learned, a study of policies that use incentives to fill a shortage seem less compelling. After all, the operation of such programs is straightforward. Offering a bonus for certain types of credentials is more likely to attract teachers with those credentials than not offering the bonus. Only one question remains: How responsive are teachers to bonuses? Performance-based pay, on the other hand, is a subject that has been studied a great deal, but the important impact questions have not been adequately addressed to date. Rewarding skills and

² An exception would be if program designers claimed that the mere existence of a plan motivates teachers despite the lack of pay differentiation. Such claims are not easily found in the existing literature on teacher pay reform.
knowledge is an approach that has attracted a lot of curiosity from researchers, but research (cited in Chapter 2) already exists and more is underway to study the validity of teacher assessments and their relationship to classroom performance.

Our application of the criteria to the four broad intervention types does not yield a clear favorite. Table III.1 shows how an attempt to rank the options according to the above criteria can be used to justify focusing the study on any one or a combination of approaches to teacher pay reform. Other factors, however, may help tip the balance in favor of one approach or the other. Once this process leads to decisions about which approach or approaches to include in a study, the research questions can be revised and tailored to the problem at hand.

<table>
<thead>
<tr>
<th>What is rewarded?</th>
<th>Relevance</th>
<th>Feasibility</th>
<th>Research Base</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. More work</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2. Skills and knowledge</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>3. Filling a need</td>
<td>4</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>4. Performance</td>
<td>1</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

NOTE: Ranking of 1 means first choice on the criterion list; ranking of 4 means last choice.

C. MECHANISMS FOR IDENTIFYING STUDY SUBJECTS

Another way to narrow the possibilities for study is to consider the practical considerations about how to mount the intervention. Experience with successful (viable) pay reform policies suggests that teachers who are subject to the compensation system should have an active role of shaping it. In terms of a research study, it suggests that the study's subjects may have to play a
large role in shaping the intervention. Thus even through ED may set some broad parameters and choose one or a combination of the four types of reward systems, many of the details about what is implemented in the study site or sites will depend on the contracting mechanism used by the study’s sponsor and the arrangements by which the intervention will be financed. The possible approaches can be grouped into three options:

1. **Conduct a federal demonstration.** A classical demonstration evaluation could be conducted in much the same way that other federal program evaluations are performed. ED would solicit grant proposals or contract bids for teacher pay reforms from the field. The solicitation could be conducted as an open-ended design competition or a tightly specified request. Either way, ED retains maximum control over the intervention selection process and is limited only by the extent of interest from the field. The major drawback of this approach is that a federally funded study would probably require not only evaluation funds but also funds for the intervention itself. In the case of most teacher pay plans, except plans executed on a very small scale, the cost is likely to be high. The most feasible strategy for a study in the context of a federal demonstration would be some form of limited planned variation.

2. **Piggyback on state or local demonstration.** Another option is to identify locations where teacher pay reform is more or less underway or planned (or mandated to be designed under a state law) and recruit the site to participate in a rigorous study sponsored by ED. This option might be the only feasible way to perform a study of a complete pay reform package. It leaves the development and implementation in the hands of those who would normally implement the policy but allows for federal access through a research partnership. Given that states, localities, or privately funded schools (such as TAP-supported schools) would probably want to document their progress, a partnership of this type with federal evaluation dollars and expertise might be welcomed.

3. **Conduct observational research.** The possibilities for learning much about the teacher pay reforms by using extant data are slim but worth considering as the costs would be dramatically lower than a prospective field study. On the other hand, observational research requires heroic assumptions in order to conclude that differences in outcomes between implementers and nonimplementers of a policy can be attributed to the policy.

The above choices can be revisited once the design options and feasibility issues are considered.
IV. DESIGN OPTIONS FOR A RIGOROUS EVALUATION

Selecting the most appropriate methods for studying alternative teacher compensation would be a relatively straightforward exercise once the research question is clearly stated, but simply deciding whether to conduct a randomized experiment or a quasi-experiment or some alternative is only one part of specifying what must be a comprehensive research design. The approach taken here sketches out a research design by posing a series of multiple choice questions ("design options") whose answers suggest how the study will be carried out. Only by considering the answers to these questions can policymakers be assured of a feasible evaluation plan.

Figure IV.1 presents a list of design options. The recommended sequence calls for first defining the object of study, that is, the treatment and counterfactual against which the treatment is to be compared. The next step is to think about the outcomes that need to be measured, the relevant units whose treatment status will vary, and, finally, the mode of treatment assignment (choice of experimental or quasi-experimental methods) that will be the centerpiece of the design. With these choices made, the other design issues may be easier to resolve. The rest of this chapter discusses the tradeoffs that must be considered in selecting design options and offers advice on each.³

³ While a randomized experiment may or may not be appropriate, the language of scientific experimentation (e.g., "treatment" and "control") is used throughout the discussion to show the connection between the research design and the principles of scientific method.

31
A. COUNTERFACTUAL OF INTEREST

A fundamental question motivating the evaluation design should ask, What comparisons do we want to make? The question defines the "treatment"—discussed in Chapter 3—and the condition to which it is compared, a counterfactual condition of interest. For a given intervention or treatment condition, measurement of the outcomes (the "factual" state) is straightforward. However, to understand the causal relationship between the treatment and the observed outcomes, we need to compare that relationship to what would have happened under
some alternative policy or intervention strategy (the "counterfactual" state). The choice of a preferred counterfactual affects the *feasibility* of approximating it by using experimental or quasi-experimental methods, but it also affects the *relevance* of the findings. We seek a counterfactual that is both relevant and feasible. A simplified list of options for the treatment and counterfactual includes the following:

1. **Strategy A versus fixed salary schedule.** In this option, the control condition would be the system of teacher compensation that prevails in the absence of reform. Given that the fixed salary schedule is so widespread in public education, the counterfactual will be fairly well defined regardless of which districts participate in the study.

2. **Strategy A versus Strategy B.** A different contrast would result if we assumed that some type of teacher pay reform would be enacted and then compared two or more alternative packages that reformed teacher pay in different ways. Such an approach is sometimes referred to as a "horse race" design. To be useful, Strategies A and B would be chosen to highlight some meaningful philosophic difference. For example, Strategy A might reward teachers based on their students’ test scores, whereas Strategy B might reward teachers based on supervisor ratings.

3. **Multifactor experiment.** A different approach would use planned variation with continuous or multilevel factors. For example, to test the effect of incentives on behavior, we could vary the size of teacher bonuses or the relative weights of different rating factors in determining the eligibility for increased compensation.

Option 1 offers several advantages over the others. It represents a common choice that future policymakers would have to weigh; the control condition is fairly well known; and it is a "high-contrast" design, implying a greater ability to detect statistically significant impacts for a given sample size. Option 1, however, may pose a practical challenge to the evaluation because it defines a treatment group that receives an intervention (including possible bonuses for teachers) and a control group that receives nothing extra. For a random assignment study of the more ambitious teach pay reforms, all stakeholders would have to be actively involved in adopting, and potentially designing the new system. This was the implication of many of the studies, cited in Chapter 2, of why teacher pay plans succeed or fail. Teachers would go through a potentially divisive process like this just to reach the stage of being eligible to be randomly
assigned. If they are then assigned to a control group, they might feel bitter and be unwilling to comply with data collection. Another risk is that, having invested effort in reaching the point of willingness to adopt performance-based pay, they might seek alternative funding sources and institute a pay reform anyway. This phenomenon is similar to what happened in two random assignment evaluations of whole school reforms in Chicago (Cook et al. 2000) and Prince George’s County, Maryland. (Cook et al. 1999).

An advantage of Option 2—seeking out alternatives or refusing to comply with study procedures—is that everyone receives some type of treatment so that the risk of perverse behavior is minimized.

Option 3 provides an opportunity to address a more tightly focused research question by varying aspects of the intervention over a well-defined range. Depending on the factors that are varied, Option 3 may require a larger sample, but it may also be more feasible to treat individual teachers or teams of teachers as the unit of intervention. Intervening at the teacher level has advantages such as lower cost and higher precision, and drawbacks, such as the risk of contaminating of the control group and the inability to replicate a full blown pay reform policy. The issue of units of intervention is discussed below.

The three design options for selecting the counterfactual are not mutually exclusive. Study subjects can be randomly assigned to different strategies and then to different variations within a strategy, allowing the evaluation to test different hypotheses. Building in multiple hypothesis tests can, however, also add to the study’s requirements and hence cost. The recommended counterfactual depends on ED’s priorities for the research questions to be addressed.

B. MEASURING OUTCOMES

Deciding what outcomes to measure and how to measure them is a necessary step in any study of teacher pay reform. The first question of what outcomes to measure should depend on
type of intervention—in this case, selection(s) from the choices arrayed in Figure I.1. The most immediate outcomes of differential pay plans are that they might change the composition of the teaching force via recruiting and retention effects and that they might change the effort that teachers devote to achieving highlighted goals via changes in the school climate. School climate effects represent whatever happens when new incentives are introduced. Some incentives cause teachers to focus on particular subjects, particular students, or particular activities over others. Some incentives make it more likely for teachers to work together while others make it less likely. Some incentives may prove highly motivating for teachers near the threshold and particularly discouraging or demoralizing for teachers who have no hope or ability to meet the standards or those who face a great deal of uncertainty about their compensation.

Teacher composition effects and productivity effects are beneficial if they translate into a rise in overall performance of the teaching staff. Some concept of “overall” teacher performance would capture all the effects previously described, but it may prove difficult to measure objectively. Methods for measuring overall teacher performance are discussed below.

Finally, there is the effect of teacher pay plans on the end users, parents and students. Final outcomes are highly relevant to policymakers and stakeholders but pose a challenge for use in an evaluation because of the difficulty of specifically attributing changes to compensation policy. Random assignment of students to treatment conditions helps a great deal, but, at the final stage, the impacts may be small relative to the many other influences such as family background and student motivation. As a result, both random assignment and large samples would be necessary to detect meaningful impacts of pay reform when final outcomes are the basis for evaluation.

---

4 Depending on the scale of the reform strategy, some effects may ripple beyond the teachers eligible for bonuses or wage increases and affect staff at neighboring schools or school districts. Accordingly, a comprehensive evaluation should be wary of unintended consequences of any targeted teacher pay reform policy.
The different (nonmutually exclusive) outcomes can be incorporated into a study in the following ways:

1. **Recruiting as an outcome.** The unit of analysis could be empty slots; the way to measure outcomes would be to determine if an empty slot was filled, when it was filled (before the school year or in what month if during the school year), and what type of teacher filled it (substitute or emergency, adequate recruit who meets the specifications, high-quality recruit who exceeds the specifications) (see Table IV.1 for a prototype of a table that could be used in an evaluation). Another approach to recruiting as an outcome would be to consider acceptable applicants as the unit of analysis. For each recruit, the impact would be the difference in acceptance rates between those subject to the alternative pay system and those not subject to the alternative pay system. Binary outcome methods, such as logit regression, are available for using covariates and parametric modeling to estimate impacts.

2. **Retention as an outcome.** Retention could be easily measured as the fraction of teachers who remain in their position for the entire school year and the fraction of teachers who remain in their position from year to year. Overall retention is not always the best indicator of policy success. An effective policy would have to result in good teachers staying more often and weaker teachers leaving more often. Evaluating such effectiveness requires knowledge of strong and weak teachers. If we assume for the moment that it would be possible to group the teachers into high and low categories, we could examine the retention rates of both groups, perhaps compared with their counterparts in a control setting, such as a school that did not participate in teacher pay reform. Figure IV.2 shows an example of this for a successful intervention. The lines on the graph are called survival curves because they represent the proportion of an initial cohort of teachers that is still teaching at each time point. (The term comes from biostatistics, where fully worked-out methods for analyzing duration data can be applied.) In the hypothetical example shown in Figure IV.2, the high-quality teachers left the school at lower rates than in the control schools, but the low-quality teachers left the school at higher rates than in the control schools. This desirable result produces a stronger pool of teachers in the treatment school (assuming replacement teachers are no worse than the leavers). The graphical representation also shows the extent to which the incentive program causes retention rates to differ by the quality of teacher. For control schools in the example, the lines (dashed) are close together and nearly parallel. For the treatment schools, the lines (solid) are some distance apart.

3. **Teacher performance: supervisor ratings.** Teacher performance is conceptually at the center of the pay reform problem, yet operationalizing it poses a serious challenge. Peer or supervisor ratings based on evaluation practices such as

---

5 The unit of intervention may have to be schools, because principals' decisions about filling slots with bonuses attached are interrelated with decisions about filling other slots and redeploying existing staff.
classroom observation have been the predominant mode for measuring teacher performance in the history of teacher pay reform (see Chapter 2). For that reason alone, it might be useful to measure supervisor ratings.

4. **Teacher performance: value added.** A highly salient measure of a program’s effectiveness is whether it makes teaching improvements that result in more learning. The effect of teaching on learning is sometimes known as value added and is typically measured by examining student achievement gains, making some adjustments for differences in student characteristics that are outside the teacher’s direct control.

| TABLE IV.1 |
| IMPACTS ON RECRUITMENT: HYPOTHETICAL EXAMPLE |

<table>
<thead>
<tr>
<th>Recruitment Outcome</th>
<th>Control Slots (No Bonus)</th>
<th>Treatment Slots (No Bonus)</th>
<th>Impact</th>
<th>p-value&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>When Position Filled</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Before school</td>
<td>40%</td>
<td>50%</td>
<td>+10%</td>
<td></td>
</tr>
<tr>
<td>First month</td>
<td>30</td>
<td>35</td>
<td>+5</td>
<td></td>
</tr>
<tr>
<td>After first month</td>
<td>10</td>
<td>5</td>
<td>-5</td>
<td>.001***</td>
</tr>
<tr>
<td>Never</td>
<td>20</td>
<td>10</td>
<td>+10</td>
<td></td>
</tr>
<tr>
<td>Type of Teacher Filling Position</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>10</td>
<td>20</td>
<td>+10</td>
<td></td>
</tr>
<tr>
<td>Adequate</td>
<td>50</td>
<td>60</td>
<td>+10</td>
<td>.001***</td>
</tr>
<tr>
<td>Below/standard/nobody</td>
<td>40</td>
<td>20</td>
<td>-20</td>
<td></td>
</tr>
<tr>
<td>Sample Size (no. of slots)</td>
<td>200</td>
<td>200</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**SOURCE:** Hypothetical data for illustration only.

<sup>a</sup>p-value is based on hypothesis test that the treatment and control group distributions are equal.
We are interested in performance in the areas targeted by the rewards system. To the extent that the measured performance is a proxy for “true performance,” the proposed evaluation would have to include some enhanced measurement that would not otherwise be feasible in an ongoing compensation system.

Measuring teacher performance for an impact evaluation is not straightforward. It is not only the most challenging aspect of the evaluation, it is the central and most challenging component of any alternative teacher pay plan. In one sense, teacher performance must be measured in order to be rewarded and therefore is measured as part of the intervention anyway. However, performance measures used in the context of an ongoing award system may be approximate and may differ for different teachers in the study (because “Treatment A” might call for teacher performance to be measured in one way, “Treatment B” in another way, and the “control” condition would not necessarily have any formal mechanism for measuring teacher performance at all).
Further, to test for unintended consequences, a rigorous study needs to measure outcomes in other areas in addition to those measured by the teacher pay system. Many alternative pay systems are hypothesized to have perverse incentives that divert resources and narrow accomplishment of the standard. For example, if the aim is to reward mathematics and reading, then history and science scores may suffer as a result of the intervention as teachers spend more time on the high-stakes subjects. The same principle applies to different types of students. If a passing standard is set at one threshold point, then we might expect teachers to expend the greatest effort in working with students nearest the threshold and pay less attention to helping those far below or above the threshold.

The recommended approach depends on the intervention under study. If the intervention is mainly targeted toward composition effects and allows for meaningful (even if gross) distinctions to be easily made between better and worse teachers, then a focus on recruiting or retention outcomes alone is practical and wise. If, however, the aim is to understand the effect of teacher incentives on teacher effort, behavior, and productivity, then it is appropriate to use multiple outcome measures to make sure that both intended and unintended consequences of pay reform policies can be measured. If the intervention is expected to have any impacts on student achievement, then student achievement should be included as a major outcome despite the challenges inherent in using student achievement data. With annual testing in every grade becoming more routine in American schools, administering new student achievement tests as part of the study may be unnecessary.

C. UNITS OF INTERVENTION

A major yet practical decision that affects the rest of the study design is the unit at which to intervene to modify the teacher pay system. We present four options and discuss the tradeoffs:
1. *Individual teachers or teaching slots.* The evaluator can allow teachers to participate in different compensation plans, each with its own set of rules for evaluation and compensation. Alternatively, the evaluator can treat teaching vacancies as the unit to which an intervention such as a signing bonus can be applied. Teacher-specific interventions make it practical to build a large sample, but some types of interventions would have spillover effects, and some types of outcomes such as school climate cannot be easily captured in this type of design.

2. *Teams of teachers.* The evaluator can identify the smallest unit that works together in a self-contained way, such as teachers in the same grade, teachers teaching the same set of children (different subjects), teachers teaching the same subject, or the entire school.

3. *Schools.* Use of schools as units of intervention represents the broadest interpretation of teams. It includes everyone in the building. The distinction from a strict team-based design is that treatment status would not vary within schools. The evaluation could focus on a single grade level or subject area (such as high school mathematics) and simply exclude the rest of the school from the study.

4. *Districts.* Districts are the most natural units in that they are often the bargaining units that negotiate teacher contracts. The challenge, though, lies in the difficulty of finding reasonable matches or a sufficiently large sample such that chance differences would average out through random assignment. A district-level analysis with a small number of districts may be unsatisfying because, even with careful matching, any treatment-control differences would be confounded with some differences that are bound to remain.

Some performance pay plans provide rewards at several levels and award individual and group bonuses. Most pay plans that target specific areas of subject shortage or overall teacher shortages in certain types of schools would be amenable to analyses with teaching slots as the unit. If the research question is posed at the most general level—what is the total effect of a pay-for-performance program?—then it may be necessary to use entire districts as the unit. When the unit issue is resolved and the significant outcomes of interest are known, then it is appropriate to conduct a statistical power analysis to determine whether the sample size requirements lead to a feasible overall design. Such analyses are useful for gauging whether a given design is feasible. In the meantime, the recommended approach is to consider experiments with schools or grades within schools as the unit of intervention when the research question is about performance-based
pay or pay for extra work. Pay for skills and pay for filling a need can be evaluated in many circumstances by using individual teachers as the unit of analysis.

D. MODE OF TREATMENT ASSIGNMENT

A fundamental design decision is what method will be used to approximate the counterfactual condition.

1. Experimental design (random assignment).

The most desirable design from the perspective of scientific method is to use random assignment within a controlled setting. Random assignment would be used to construct a control group of teachers, schools, or districts as necessary. Then differences in average recruitment, retention, school climate, student performance, and so forth between intervention and control settings could be attributed to the intervention.

The two greatest feasibility challenges to a randomized design of a study of alternative teacher pay are the difficulty in recruiting study subjects to participate and, depending on the type of units selected (see above), the cost of gathering a large enough sample to rule out chance differences between treatment and control conditions. Recruitment is always a challenge in randomized trials. The size of the challenge in the current context depends on how the treatment and counterfactual are defined and the way the policies are financed. If the evaluator is coming to the table with a generous package of bonuses and does not impose much of a burden and requires only limited paperwork to determine eligibility for bonuses, then cooperation will be not pose a challenge. On the other hand, if treatment schools are asked to undergo a risky overhaul of their compensation system and control schools are asked to accept a bonus program and then receive nothing, they may protest and refuse to participate.

On the sample size issue, if the units are districts or entire educational markets, then it might be prohibitively expensive to obtain a large number of sample members, especially as the fixed
costs per district (bargaining unit) for negotiation of cooperation may be exorbitant. On the other hand, an experiment that could be limited to individual schools, teachers, or teaching slots would be easier to mount because larger numbers of teachers could be recruited from within the same district or school.

2. Quasi-experimental designs.

In field settings, random assignment is sometimes considered impossible when, for example, subjects cannot be induced to participate or comply; such cases require the consideration of alternatives.

a. Matched comparison design (prospective). A reasonable alternative to random assignment would be to identify subjects (teachers, schools, districts) that are implementing alternative pay schemes and seek out subjects that appear to be identical in every way except for pay policy. The major drawback with matched comparison designs is that the evaluator can never be certain the match has accounted for all unobserved differences. In terms of costs, a matched comparison design that requires the recruitment and following of treatment and comparison sites is probably just as costly as a randomized experiment.

b. Instrumental variables, regression discontinuity. Natural experiments sometimes present themselves to researchers and provide opportunities to learn about policies without intervening in costly and intrusive ways. For example, recruitment rates could be examined for teacher position openings in a range of schools defined by different levels of poverty. If there is a break in recruitment rates at the threshold where a school qualifies as "high-need" and therefore is eligible for state-funded signing bonuses, then the size of the discontinuity can be attributed to the bonus policy. Such natural experiments provide a tempting low-cost opportunity to learn about teacher pay reform but cannot be crafted to suit any research question, and should be undertaken only when the opportunity coincides with the research question.

c. Interrupted (school) time-series analysis. Some have suggested the use of school-level time-series data to identify impacts of education policies on student achievement (Bloom 2003). The idea is to measure four or five cohorts’ test scores before the introduction of an intervention and then estimate program impacts as deviations in each successive year from the preprogram time trend. These methods have received some criticism because they rely too much on the stability of achievement trends across cohorts within a school. Glazerman (2003) shows how the bias in such cohort difference measures is compounded by initial cohort differences and
differences in schooling experiences throughout the cohort members' schooling careers. A preferable method would rely on longitudinal student data in which deviations from a preprogram trend would be more intrinsically meaningful and attributable to educational interventions.

3. Alternatives to the impact evaluation paradigm.

Causal knowledge about the effect of a program initiative is not always the goal of research on teacher pay reform. Many interesting questions can be answered without estimating a counterfactual. One example is the quality of teacher assessments used in performance pay systems to determine eligibility for incentive awards.

A valuable contribution to the field could be made by measuring teacher productivity with a highly sophisticated and in-depth but impractical tool and using that productivity measure as a benchmark to assess the quality and validity of more practical tools that are used to measure teacher effectiveness in ongoing teacher pay reform policies. Examples of such imperfect but practical tools include the National Board certificate, supervisor ratings, and simple test score gains.

The recommendation for which design type to adopt depends critically on the answer to the earlier questions about the nature of the research question, the type of treatment/counterfactual contrast, the outcomes of interest, and, most important, the units of intervention. Given how little is known about why teachers participate in incentive programs and why schools or districts adopt them, a randomized experiment is strongly preferred over quasi-experimental designs. To make the experiment feasible may require scaling back the ambitiousness of the research question to create an attainable but still useful research goal. Chapter 5 presents some examples.
V. PUTTING IT ALL TOGETHER—ILLUSTRATIVE EXAMPLES OF EVALUATION DESIGNS

The method for constructing a study design followed in Chapter 4 is useful for narrowing options. However, all of the individual decisions are interdependent. Ultimately, the pieces must be assembled into a coherent strategy for evaluating alternative pay plans. The following study designs are illustrative examples of how to combine the elements discussed above. The examples are meant to prompt discussion and consideration of the practical consequences of selected options, not to limit the possibilities. Other design templates can be developed and refined as ED narrows its priorities.

A. DESIGN A. CLASSICAL DEMONSTRATION PROGRAM EVALUATION (RANDOMIZED EXPERIMENT)

Research question: What is the impact of the Teacher Advancement Program on student achievement and other outcomes?

Unit of analysis: Schools

Outcomes: Value added in academic subjects targeted by the performance measurement system, value added in nontargeted subjects, teacher assessments, retention, school climate (attitudes, collegiality)

Treatments, counterfactual: TAP, alternative incentive program, uniform salary schedule

Treatment assignment: Randomized trial

The study design would test the relative effects of a packaged pay reform program, for example, the Teacher Advancement Program, on student achievement and other outcomes. The TAP would be compared to both an alternative policy, in this case a system such as the Douglas, Colorado, pay-for-performance plan that emphasizes pay for knowledge and skills, and a traditional fixed salary schedule.
The design would call for identifying schools whose staff would be willing to participate in the study, grouping them into triplets that are as nearly identical as possible but are situated in different educational markets, and randomly assigning each to one of the three treatment conditions. A study of this type could require at least 15 to 21 schools and perhaps many more.

B. DESIGN B. PLANNED VARIATION EXPERIMENT

Research question: How responsive are different types of in-demand prospective teachers to hiring incentives?

Unit of analysis: Schools/position openings

Outcomes: Recruitment, retention, administrative costs

Treatments: Recruitment incentives of varying amounts, terms, forms (cash versus loan repayment or housing assistance)

Treatment assignment: Random assignment

The study could be conducted by announcing a grant competition for schools that experience difficult in seeking selected categories of teachers and for high-need schools that experience difficulty in finding certified teachers. Grant applicants (school principals) who agree to participate in the study would then be randomly assigned to receive different levels and types of funding for recruitment incentive programs. Some schools would be allowed to offer prospective teachers meeting the need criteria a one-time cash payment through the study. Others would be allowed to offer bonuses in installments over time, and still others would be allowed to offer only loan repayment or housing assistance. Furthermore, the researcher would systematically vary the amounts that schools would be allowed to disburse. In general, schools receiving funding would not be told about the incentive packages assigned to other schools in the study. The evaluator would measure recruiting and retention outcomes and monitor the administrative costs of operating the incentive programs. The study would produce, among other results, estimates of the response curve that shows how variation in the bonus amount increases
the probability of a successful hire before the start of the school year. The shape of the curve might differ for teachers in different subject areas and for schools with higher or lower levels of poverty or disadvantage.

C. DESIGN C: VALIDITY AND COST OF TEACHER PERFORMANCE MEASUREMENT

Research question: How accurate and reliable are low- and medium-cost measures of teacher performance?

Unit of analysis: Teacher/assessment instrument

Outcomes: Correlation of teacher performance indicator with benchmark indicator

Treatment: Superficial and intensive measurement of teacher performance, using classroom observation, portfolio measures, student test scores

Treatment assignment: N/A. Comparisons between different measures, not different individuals

The study would not be a traditional impact evaluation. Rather, it would be a precursor study to determine if an accurate performance measurement of the type needed to make a merit pay program function is possible. In a school district using performance-based pay, the researchers would draw a random sample of teachers and make intensive measurements of their performance, including highly sophisticated measures of value added (perhaps with random assignment of children to classrooms). Less intensive measures would be implemented, including knowledge and skills assessments and performance assessments, and principals would be asked to predict the value added by their teachers. Each performance measure, including principals' subjective ratings, would be compared to the gold standard benchmark performance measure and would be arrayed by cost to permit readers to judge each indicator's cost-effectiveness. A process analysis could be conducted to see the extent to which the measurement itself changed practice and to characterize the burden imposed on staff and students by the different types of performance measurement.
D. CONCLUSION

The three studies described above provide examples of ways to think about designing a research agenda to learn more about teacher pay reform. The studies range from highly ambitious and probably costly (Design A) to somewhat less ambitious and less costly (Designs B and C). Many more design possibilities can be developed, depending on the needs of the education policy community. In general, the harder-to-answer questions entail greater risk and cost, but, if the will exists to study a particular problem, most technical challenges can probably be overcome.
REFERENCES


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