Feasibility of a Star Teacher Demonstration

February 28, 2007

Jeffrey Max
Allison McKie
Steven Glazerman

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

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

Project Officer:
Stefanie Schmidt

Project Director:
Steven Glazerman
## CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td></td>
</tr>
<tr>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>A. BACKGROUND</td>
<td>1</td>
</tr>
<tr>
<td>B. METHODS AND ORGANIZATION OF THE REPORT</td>
<td>2</td>
</tr>
<tr>
<td>C. RESEARCH QUESTIONS AND OVERVIEW</td>
<td>2</td>
</tr>
<tr>
<td>II</td>
<td></td>
</tr>
<tr>
<td>DESIGN AND IMPLEMENTATION OF A STAR TEACHER RECRUITMENT PROGRAM</td>
<td>5</td>
</tr>
<tr>
<td>A. DEFINING AND SELECTING STAR TEACHERS</td>
<td>5</td>
</tr>
<tr>
<td>1. How Many Years of Student Learning Gains Are Needed for the Value-Added Analysis?</td>
<td>5</td>
</tr>
<tr>
<td>2. What Criteria Should the Demonstration Use Besides Value-Added Indicators?</td>
<td>7</td>
</tr>
<tr>
<td>3. Should There Be Any Exclusion Criteria for the Demonstration?</td>
<td>8</td>
</tr>
<tr>
<td>4. Should the Demonstration Have a Within- or Between-District Focus?</td>
<td>9</td>
</tr>
<tr>
<td>B. RECRUITING STAR TEACHERS WITH INCENTIVES</td>
<td>10</td>
</tr>
<tr>
<td>1. What Bonus Amount Is Needed to Attract Star Teachers?</td>
<td>10</td>
</tr>
<tr>
<td>2. What Forms of Nonmonetary Incentives Would Attract Star Teachers?</td>
<td>12</td>
</tr>
<tr>
<td>3. Should Star Teachers Be Required to Make More Than a One-Year Commitment?</td>
<td>14</td>
</tr>
<tr>
<td>C. PLACING STAR TEACHERS IN HIGH-NEED SCHOOLS</td>
<td>15</td>
</tr>
<tr>
<td>1. How Should the Demonstration Define High-Need Schools?</td>
<td>15</td>
</tr>
<tr>
<td>2. What Are the Key Aspects of the Star Teacher Placement Process?</td>
<td>18</td>
</tr>
<tr>
<td>3. Should Star Teachers Be Placed Into High-Need Schools Individually or as a Group?</td>
<td>20</td>
</tr>
<tr>
<td>D. ADDRESSING KEY STAKEHOLDERS</td>
<td>20</td>
</tr>
<tr>
<td>1. What Are the Potential Concerns of Teachers’ Unions?</td>
<td>20</td>
</tr>
<tr>
<td>2. How Will Sending Schools Respond to the Program?</td>
<td>22</td>
</tr>
<tr>
<td>3. What Are the Potential Concerns of Receiving Schools?</td>
<td>23</td>
</tr>
</tbody>
</table>
### CONTENTS (continued)

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>III</td>
<td>EVALUATION OF THE STAR TEACHER PROGRAM</td>
</tr>
<tr>
<td>A.</td>
<td>THE COUNTERFACTUAL</td>
</tr>
<tr>
<td>1.</td>
<td>What Is the Most Policy-Relevant Counterfactual?</td>
</tr>
<tr>
<td>2.</td>
<td>What Is the Most Feasible Approach to Approximating the Counterfactual?</td>
</tr>
<tr>
<td>3.</td>
<td>Implications for Evaluation Design</td>
</tr>
<tr>
<td>B.</td>
<td>UNIT OF RANDOM ASSIGNMENT</td>
</tr>
<tr>
<td>1.</td>
<td>Randomly Assign Students</td>
</tr>
<tr>
<td>2.</td>
<td>Randomly Assign Schools</td>
</tr>
<tr>
<td>3.</td>
<td>Seek More Expert Advice Before Committing to a Design</td>
</tr>
<tr>
<td>C.</td>
<td>SAMPLE SIZE AND POWER</td>
</tr>
<tr>
<td>1.</td>
<td>How Many Schools Are Required?</td>
</tr>
<tr>
<td>2.</td>
<td>Differences in Costs</td>
</tr>
<tr>
<td>3.</td>
<td>Cost-Precision Tradeoffs Illustrated</td>
</tr>
<tr>
<td>D.</td>
<td>DATA COLLECTION AND MEASUREMENT OF OUTCOMES</td>
</tr>
<tr>
<td>1.</td>
<td>What Are the Study’s Main Outcomes and How Should They Be Measured?</td>
</tr>
<tr>
<td>2.</td>
<td>What Grades and Subjects Should the Study Include?</td>
</tr>
<tr>
<td>3.</td>
<td>What Other Data Would Be Useful?</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>39</td>
</tr>
</tbody>
</table>
TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>II.1</td>
<td>DATA NEEDED TO CALCULATE STUDENT LEARNING GAINS</td>
<td>8</td>
</tr>
<tr>
<td>II.2</td>
<td>PRECEDENTS FOR THE STAR TEACHER PROGRAM</td>
<td>11</td>
</tr>
<tr>
<td>II.3</td>
<td>NUMBERS OF SCHOOLS, DISTRICTS, AND STATES POTENTIALLY AVAILABLE FOR STUDY</td>
<td>17</td>
</tr>
<tr>
<td>II.4</td>
<td>ILLUSTRATIVE PROJECT TIMELINE</td>
<td>19</td>
</tr>
<tr>
<td>III.1</td>
<td>MINIMUM NUMBER OF TEACHING SLOTS PER SCHOOL-GRADE REQUIRED BY DESIGN</td>
<td>30</td>
</tr>
<tr>
<td>III.2</td>
<td>SAMPLE SIZE REQUIREMENTS BY DESIGN</td>
<td>32</td>
</tr>
<tr>
<td>III.3</td>
<td>ILLUSTRATION OF THE COST-PRECISION TRADEOFF</td>
<td>35</td>
</tr>
</tbody>
</table>
CHAPTER I
INTRODUCTION

Research findings show that teacher quality makes a difference in student learning (Rowan et al. 2002), but the most troubled schools experience difficulty in attracting and retaining effective teachers, especially in high-demand subject areas (McCreight 2000; Curran et al. 2000). The logical conclusion of these findings is that a maldistribution of teacher quality will lead to a maldistribution of learning opportunities for students. Several states and school districts have developed incentive policies to address this problem, but the research that accompanies such programs has been lacking.

In response, the U.S. Department of Education’s Institute of Education Sciences (IES) contracted with Mathematica Policy Research, Inc. (MPR), to consider the design of new programs that can be studied rigorously and yield lessons for the field. The ultimate goal is to carry out a demonstration that would advance the state of knowledge on how high-performing teachers can be used to help high-need schools and lead to helpful policies. This report proposes options for designing such a demonstration and discusses the feasibility of the options, offering recommendations where possible. MPR seeks to recommend the key features of the intervention to be studied and an evaluation of its effectiveness.

A. BACKGROUND

In summer 2006, IES asked MPR to consider two approaches to the demonstration. One approach would identify highly effective teachers and provide them with incentives to teach in high-need schools. The evaluation component of the demonstration would be designed to estimate the impact of such teachers on student achievement in their new schools. We refer to this approach as the “star teacher recruitment” program. The other approach, which is similar but not the same, would offer financial incentives for qualified teachers to accept positions in high-demand subject areas in high-need schools. We refer to this approach as a “differential pay” experiment. The goal of the feasibility study is to flesh out the details for both approaches and assess the feasibility and cost-effectiveness of carrying out a demonstration of one or both approaches.
B. METHODS AND ORGANIZATION OF THE REPORT

Some key principles guide our effort to design the demonstration. First, the program or programs studied in the demonstration should be policy-relevant. That is, they must present options that school districts could realistically put in place. Second, they must provide a cost-effective, informative test of policy. Third, the evaluation component should be rigorous and use random assignment methods. Finally, given that feasibility is a great concern, we will consider demonstration designs that are opportunistic with respect to site selection in order to maximize the chance of success. The goal is to obtain a sample that is not idiosyncratic in a way that makes it hard to extrapolate to other school districts but that is not necessarily statistically representative of schools or school districts nationally.

This report summarizes the initial findings and recommendations of our work on the feasibility study carried out in summer and fall 2006, focusing on the first of the two proposed demonstration ideas. We identified and interviewed in detail more than 15 individuals with relevant experience or expertise with school culture, teacher recruitment and hiring, school reform, teacher compensation, school testing data, or research methods. These individuals included urban, suburban, and rural school superintendents; state education office and school district staff; teacher union representatives; policy researchers; and staff from programs such as Teach For America and The New Teacher Project. In addition, we reviewed the relevant published literature and obtained unpublished reports on programs with components similar to those under consideration for the proposed demonstration.

The rest of this chapter lists the research questions and gives an overview of the star teacher recruitment program. Chapter II discusses options and offers recommendations on the design of the star teacher recruitment program. Chapter III discusses the design of an evaluation of the star teacher recruitment program. In parallel with Chapters II and III, a future report will include Chapters IV and V, which will discuss the differential pay program, covering program design and evaluation design, respectively.

C. RESEARCH QUESTIONS AND OVERVIEW

The goal of the star teacher demonstration is to determine whether it is possible to improve struggling schools by identifying and recruiting strong teachers from more successful schools. This report defines a program that would identify and recruit such “star” teachers and outlines an evaluation that would address some specific research questions. The central research question follows:

1. What impacts do star teachers have in struggling schools? IES, who is sponsoring the study, is mainly interested in the impacts on student achievement.

The jurisdictions that were represented include: Palm Beach County, Florida; Hamilton County, Tennessee; New York City; Philadelphia, Pennsylvania; several schools and districts in Virginia; and the state departments of education in California, Virginia, and Arkansas.
In other words, can star teachers (those who have shown evidence of producing student achievement gains under favorable conditions) also generate student achievement gains in the high-need schools targeted for intervention? When asking these questions, investigators need to be clear that the impacts are measured relative to some counterfactual state, which is the set of outcomes that would have been realized in the absence of the program. We describe the counterfactual for this study more concretely in Chapter III.

In addition to the main research question, IES has an interest in several secondary research questions, such as the following:

2. What is the impact of introducing a cadre of star teachers to a high-need school? The feasibility analysis below discusses the tradeoffs involved in designing the star teacher intervention around recruitment of an individual teacher to each targeted high-need school, several teachers per school, or both approaches.

3. What is the impact of incentives on the behavior of teachers, specifically their decision to transfer or relocate to and remain in a high-need school? The answer to the question obviously depends on the nature of the incentives offered and the circumstances under which they are offered. Given that experimental variation of the incentive package is not likely to be practical for the proposed evaluation, we explore ways in which the demonstration could offer lessons about incentives based on nonexperimental evidence.

The questions that IES has posed are focused on the high-need schools themselves, not on the sending schools from which the star teachers are recruited. In a benefit-cost analysis from the perspective of the school district or state education agency, any positive impacts of the star teacher program on receiving schools could potentially be offset by negative impacts on the sending school. However, even if the net impacts (positive impacts on receiving schools plus the negative impacts on sending schools) are zero or slightly negative, policymakers may prefer the program if it reduces inequality in educational opportunities. The current proposed demonstration, therefore, focuses mainly on the receiving schools. To provide some evidence on the consequences for sending schools, a process evaluation could include some measures of the qualifications of those who replace the star teachers in the classrooms that they leave.

We envision a star teacher program as a prestigious and rewarding fellowship or sabbatical program for highly accomplished teachers to help struggling schools raise student achievement through reconstitution—either limited or wholesale. A program implementation team would carry out the following steps: (1) recruit district(s) to participate; (2) identify high-need “receiving” schools that have or can create at least one opening in targeted grade/subject areas; (3) estimate productivity (“value-added”) scores for all teachers in the district or region to identify potential star teachers; (4) publicize the program and invite star teacher candidates to apply formally; (5) select star teacher finalists; (6) arrange interviews with principals in receiving schools; (7) complete placement and transfer paperwork; and (8) disburse bonuses or other incentives. An evaluation team would observe program
implementation, conduct random assignment in receiving schools, collect data, and produce reports on the implementation and impact of the intervention.
This chapter addresses four aspects of the program’s design: (1) the star teacher selection process, (2) incentives offered to star teachers, (3) teacher placement in high-need schools, and (4) overall feasibility from the perspectives of key stakeholders. To see how these elements fit into the evaluation design, readers should see Chapter III, where we consider different counterfactual conditions, units of analysis, sample size configurations, and data collection strategies. In general terms, the evaluation design will entail random assignment of students or schools to a treatment group, where a school may fill the teaching vacancy with a star teacher, or a control group, where the teaching slot is filled in the usual manner. This chapter is concerned with identifying, recruiting, and placing the treatment teachers.

A. DEFINING AND SELECTING STAR TEACHERS

The evaluator first needs to establish a process for identifying and selecting teachers who are highly effective in contributing to student learning gains. The process should be credible and equitable in the view of diverse stakeholders. Otherwise, the demonstration could experience difficulty in obtaining buy-in from educators or district administrators who question the fairness or efficacy of the selection process. We recommend a two-step process. The first step would use student learning gains to identify candidates for outreach and recruiting. The second step would involve in-person interviews to select teacher finalists. Below, we address key implementation issues related to the teacher selection process.

1. How Many Years of Student Learning Gains Are Needed for the Value-Added Analysis?

The demonstration will identify an eligible pool of teachers in each district who have successfully demonstrated their ability to raise student achievement. “Value-added” analysis measures a teacher’s contribution to student learning gains beyond other factors outside
their control such as students’ earlier achievement or family background, and provides an efficient method for pre-screening effective teachers. Outreach and recruitment efforts will target teachers identified through this initial screening process. The evaluator has a choice in deciding how many years of student learning gains to include in the value-added analysis.

*We recommend using three years of student learning gains for the value-added analysis.* Our recommendation stems from two considerations: the amount of data needed for accurately estimating student learning gains and the number of sites with the capacity to generate those estimates.

To account for the imprecise nature of student test scores, the analysis of student learning gains requires the pooling of data from several years. Student test scores are subject to random, one-time events that can make it difficult to distinguish teachers’ true effect on student achievement based on only one year of data (Kane and Staiger 2002). Moreover, measures of student learning gains are based on a relatively small sample of students for each teacher, as teachers typically have fewer than 25 or 30 students in a class. Small sample sizes result in greater uncertainty, and thus less precision, when trying to estimate teacher contributions to student learning gains. To address such variability in student test scores, the demonstration should rely on at least three years of student learning gains per teacher.2

The evaluator should consider whether an ample number of sites have the necessary data and data system available to calculate three years of student learning gains. Thirteen states report that they maintain a data system with the basic capacity needed to generate value-added estimates of teacher performance (Data Quality Campaign 2005). The three requirements for a suitable data system include: (1) a longitudinal data system with student and teacher identifiers, (2) the ability to match student test scores from year to year, and (3) the ability to match teachers and students. In addition, some universities, including the University of Texas at Dallas, and student testing companies, such as the Northwest Evaluation Association (NWEA), maintain teacher and student test score data for states and districts.

The data demands of the value-added analysis should be balanced against the feasibility of identifying an adequate number of demonstration sites with the necessary data capacity. States or districts that have more recently developed the data-gathering capacity for value-added analysis may not have compiled enough years of student test score data in their new data systems. Calculating value-added estimates based on three years of student learning gains requires four years of test score data, assuming that testing occurs once per year (Table II.1). If the demonstration begins in fall 2008 and recruiting begins before the end of the previous school year, sites would need longitudinal data systems with student test score data

---

2 While minimal evidence suggests that a teacher’s first two years of teaching are predictive of gains in their third year (Gordon et al. 2006), this exploratory evidence does not justify the use of two years of learning gains to calculate value-added measures for teachers at all experience levels.

---

II: Star Teacher Program Design
Table II.1 Data Needed to Calculate Student Learning Gains

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fall</td>
<td>Spring</td>
<td>Fall</td>
<td>Spring</td>
<td>Fall</td>
</tr>
<tr>
<td>Two Years of Learning Gains</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Required data points</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teaching experience</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Three Years of Learning Gains</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Required data points</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Teaching experience</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

dating back to spring 2004. Table II.1 also indicates that three years of student learning gains will provide value-added estimates only for teachers with at least three years of experience in the district and in the same subject.

2. What Criteria Should the Demonstration Use Besides Value-Added Indicators?

Additional selection criteria besides value-added analysis can lend the program credibility among stakeholders who do not understand or trust value-added analysis or view such analysis as an incomplete measure of teacher quality. Moreover, while researchers increasingly use value-added analysis to measure teacher effectiveness, the appropriateness of value-added analysis in informing high-stakes decisions or teacher rankings has been questioned because of an insufficient research base (McCaffrey et al. 2003). After identifying eligible teachers based on value-added analysis, the evaluator should implement further criteria for the selection of star teachers. Any additional steps in the selection process should attempt to screen teachers who are likely to succeed in the challenging environment of a high-need school, though without unduly burdening them.

We recommend including an in-person interview as part of the selection process for star teachers.

A carefully structured in-person interview with a well-defined rubric to evaluate each candidate is an efficient and relatively inexpensive method for identifying high-quality teachers. We spoke with several experts who asserted that direct interviews offer a cost-effective way to differentiate teacher candidates. For example, one approach presents a teacher with student work that is below, on, and above grade level and asks the teacher to identify each student and describe strategies for addressing each student’s needs. Interviews would take about 60 minutes per candidate. Our discussions with the New Teacher Project and Teach for America, two programs with extensive experience in placing teachers in high-need schools, confirmed that a one-on-one interview can also be used to predict a teacher’s ability to adapt to and persist in the environment of a high-need school.
We believe that other options for screening teachers would be either too burdensome for teacher candidates or too costly for the program. For example, classroom observations of all teacher candidates would be expensive and require substantial time and effort. A teacher’s union representative whom we spoke with was critical of classroom observations, stating that they primarily offer information on teacher routines without providing an opportunity to probe how teachers think. Other programs, such as Teach for America and the New Teacher Project’s Oakland Teacher Corps and New York Urban Teachers have used writing assignments, mock lessons, or group performance events in addition to interviews, but they involve a tradeoff between imposing more selective criteria and raising teacher and program costs of applying to the program. Excessive or burdensome entry requirements may discourage teachers from applying for a star teacher position. While principal or peer recommendations may be informative, principals or teachers at the sending school have a disincentive to provide a recommendation that could cause their school to lose a highly effective teacher.

3. Should There Be Any Exclusion Criteria for the Demonstration?

Teacher eligibility criteria must be clearly defined in a way that does not produce unintended negative consequences for the demonstration. The star teacher program is not intended to remove highly effective teachers from high-need schools. A basic assumption undergirding the program is that effective teachers are not found, or at least are underrepresented, in high-need schools. Therefore, the goal of the program is to address teacher maldistribution. The demonstration can most directly address this goal by targeting the star teacher program to teachers in low-need schools and excluding those who already teach in high-need schools. However, fairness dictates that teachers serving the most disadvantaged students should not be excluded. Accordingly, we recommend use of a school’s Adequate Yearly Progress (AYP) as both a school eligibility criterion (discussed below) and a teacher exclusion criterion.

We recommend that teachers whose schools have not made AYP in some set of recent years (to be determined) should be deemed ineligible for the incentive package offered through the star teacher program.

Excluding teachers in low-performing schools accomplishes two purposes. First, it ensures that the star teacher program does not adversely affect low-performing schools by encouraging the transfer of their high-performing teachers. Second, it satisfies a goal stated by IES for the proposed demonstration to attract new talent, rather than reward teachers who have already made the decision to teach in high-need schools.

Before proceeding with the demonstration, IES should consider whether teachers in high-need schools will object to being excluded from the value-added analysis. Teachers may feel that the demonstration unfairly assumes there are no highly effective teachers in low-achieving schools. Additionally, while useful for research, the program’s focus on recruitment of high quality teachers from low-need schools rather than retention in high-need schools may hurt the high-need school’s ability to hire and develop “raw talent.” Teachers would face incentives to prove themselves eligible in other schools before ever agreeing to teach in the high-need school.
An alternative would be to offer retention incentives for teachers who meet the eligibility standard but already work in high-need schools. Teacher transfer incentive programs that we identified around the country, such as the Arkansas High Priority District Bonus Program and the Virginia Teacher Incentives for Hard-to-Staff Schools program, do just that. The disadvantages to adding a retention incentive, however, are that it would be more difficult for the evaluation to isolate the source of any treatment effects and that the demonstration costs could rise unpredictably, as the program incurs obligations to pay all eligible teachers in the demonstration schools. Furthermore, if we randomly assign schools (see Chapter III), then excluding control school teachers from consideration would be necessary to preserve the integrity of the experiment.

4. Should the Demonstration Have a Within- or Between-District Focus?

The star teacher program can generate a larger pool of talent from which to identify eligible teachers by targeting teachers in a whole region surrounding the high-need schools, not just in the same school district. Interdistrict transfers allow the possibility of addressing the maldistribution of teacher quality across districts. However, such transfers raise several implementation challenges that heavily outweigh their benefits.

*We recommend a focus on within-district teacher transfers.* We address each of the challenges associated with across-district transfers below, including: (1) the need for a separate hiring process, (2) the transfer of employment between districts, and (3) the relocation costs of transfers across districts.

Ensuring the timely placement of out-of-district teachers in high-need schools would be a challenge and may require a separate hiring process. The process used by districts to hire new teachers presents two problems, especially in large urban districts. First, the process is often delayed by the teacher transfer period, late vacancy notification requirements, and late budget timelines (Levin and Quinn 2003). As a result, the demonstration may not place out-of-district star teachers in vacancies until late in the summer, which is when most hiring occurs for urban districts (Levin et al. 2005). Qualified teachers, even those interested in high-need schools, are typically not willing to wait until late summer for their teaching placement (Levin and Quinn 2003). Second, the voluntary transfer process may restrict principals’ choices for filling vacancies. Given that the voluntary transfer process occurs before the teacher hiring process, voluntary or involuntary transfers may fill existing vacancies. One way to circumvent these challenges would be to negotiate an agreement with the union that allows the demonstration to use a separate hiring process to fill vacancies in the participating high-need schools. Restricting the program to within-district transfers would allow the demonstration to rely on the voluntary teacher transfer process for placing star teachers.

Challenges arise when star teachers change employers to become employees of the study district. Translating teachers’ years of experience into the district’s pay scale may be difficult or beyond the program’s control. Depending on the comparability of teacher pay scales, teachers from other districts may experience a decline in salary when they transfer to the study district. Even though many teachers are covered by statewide retirement systems,
some large districts have their own retirement systems that may or may not have reciprocal agreements with the state system (Ruppert 2001). The demonstration may be able to establish agreements with sending districts to allow star teachers to take a leave of absence for participation in the program. Kentucky’s Highly Skilled Educator Program follows such an approach and enters into an agreement with sending districts to continue paying teachers’ salary and benefits while teachers are on sabbatical and participate in the program. Nevertheless, coordinating retirement, health, and other fringe benefits, as well as salary differences could be a formidable barrier to implementation across districts.

The transfer of teachers across districts also poses a challenge in terms of relocation costs for both the program and the star teachers. A transfer from one district to another to join the program would potentially uproot teachers, perhaps triggering the need to relocate and/or enroll in a new health benefit system while forgoing the guarantee of employment with their sending district. Even though within-district transfers remain employees of the district and have the option to transfer to a different school voluntarily, teachers from outside the district would have to reapply for a position in their original district if they wished to return after a fixed number of years. Compensating teachers for these additional costs would be expensive, but failure to do so would discourage the participation of teachers from outside the district.

Our recommendation to focus on within-district transfers avoids these challenges by using the voluntary teacher transfer process to place star teachers and allowing teachers to remain employees of the district, where few teachers will need to relocate. One drawback to our recommended approach is that the demonstration would need to be implemented in a large, diverse district with sufficient variation in value-added teacher effectiveness. While studies of large, urban districts (i.e., Chicago, Los Angeles, and two districts in New Jersey) find evidence of substantial variations in teacher value-added measures, the demonstration may have to exclude small or medium-sized districts with substantially less variation (Aaronson et al. 2003; Gordon et al. 2006; Rockoff 2004). This limitation can be addressed through careful district selection and recruitment.

B. RECRUITING STAR TEACHERS WITH INCENTIVES

Encouraging effective teachers to leave their current placement and transfer to a high-need school is the most challenging and risky aspect of the demonstration. It requires a carefully structured set of incentives. In this section, we cover three aspects of the incentive package: (1) the amount of the star teacher award, (2) the use of nonmonetary incentives, and (3) the length of the teaching commitment.

1. What Bonus Amount is Needed to Attract Star Teachers?

The demonstration should establish an incentive amount that is sufficient to attract an adequate pool of highly effective teachers at a realistic cost to the evaluation. The use of a monetary incentive to attract teachers is supported by evidence that teachers respond to salaries (Allen 2005) and that teacher mobility decisions are related to relative salary differences between districts (Hanushek et al. 2001; Imazeki 2005). Yet, the research
provides minimal guidance on the size of the bonus that attracts teachers. To recommend an incentive amount in line with current practice, we combine information on four programs similar to the star teacher program with information gathered from interviews with education officials.

Contingent on further input from a panel of experts, we recommend a bonus worth at least 25 of the average teacher salary in each site. For example, if the average annual salary is $55,000, the bonus would be at least $13,750. The amounts that should be paid up front versus the amount that is paid for completion of one, two or more years of service should be determined after further consultation.

The experience of four programs that resemble the star teacher program suggests that the incentive should be at least 25 percent of teacher salary. The programs offered an average incentive equal to 20 percent of the average teacher salary, with amounts ranging widely from 9 to 30 percent (Table II.2). Virginia’s program offered the largest award, at 30 percent, but also established more stringent eligibility requirements than other programs. Although California’s program set a much lower incentive amount at 9 percent, a state official overseeing the program described the incentive as inadequate and recommended a bonus more in line with Virginia’s incentive, recommending 35 percent. The 18 percent incentive offered in Mobile County did not attract enough teachers to fill vacancies in the five reconstituted schools. Rather than increase the incentive amount, the district lowered eligibility criteria to attract more applicants. Palm Beach County’s incentive, worth 21 percent of the average salary, received mixed reviews from the two district officials whom we interviewed.

Table II.2 Precedents for the Star Teacher Program

<table>
<thead>
<tr>
<th>Location</th>
<th>Program Name</th>
<th>Annual Award as Percent of Average Teacher Salary</th>
<th>Teacher Eligibility Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Palm Beach County, FL</td>
<td>Teacher Transfer Bonus Program</td>
<td>21%</td>
<td>Value-added analysis</td>
</tr>
<tr>
<td>Mobile County, AL</td>
<td>Transformation Schools</td>
<td>18%</td>
<td>Years of experience</td>
</tr>
<tr>
<td>California</td>
<td>NBPTS Certification Incentive Program</td>
<td>9%</td>
<td>NBPTS certification</td>
</tr>
<tr>
<td>Virginia</td>
<td>Incentive Program to Attract and Retain Teachers in Hard-to-Staff Schools</td>
<td>30%&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Evidence of student performance, years of experience, national certification, performance evaluations, letters of recommendation, professional leadership</td>
</tr>
</tbody>
</table>

<sup>a</sup>Virginia offered a bonus worth 35 percent of the average teacher salary in 2004–2005 and 30 percent in 2005–2006.
Our discussions with education officials and reading of the available research evidence also point toward an incentive of at least 25 percent. Hanushek et al. (2001) estimated that an additional 20 to 50 percent in salary is needed to improve teacher retention in schools with high proportions of low-performing and minority students. A union official in Denver who helped design that city’s pay-for-performance program concluded that teachers would need a salary increase of about 20 percent to encourage them to relocate or transfer. The teachers in Denver reported in focus groups and interviews that they would be willing to transfer for a 5 to 10 percent increase, although their willingness to transfer was contingent on a range of favorable working conditions. Based on the combined evidence from existing programs and education officials, we estimate on a preliminary basis that the bonus amount needed to attract teachers to high-need schools would need to be at least 25 percent.

The star teacher bonus award can be structured both to attract teachers to the program and to offer an incentive for retention. A one-time bonus can be designed with recruitment and retention features if it is paid out at the end of the school year or levies a repayment penalty prorated for the amount of time completed. Alternatively, the bonus can be paid out in installments, with a partial payment made at the outset of the teaching assignment and the remainder paid out at later points. For example, Palm Beach County and Mobile County offered 50 to 75 percent of their bonus as an upfront payment and the remaining amount at the end of the year. Both counties conditioned the final payment on performance measures. Given the challenge associated with attracting and retaining star teachers, we recommend structuring the incentive to encourage both recruitment and retention.

2. What Forms of Nonmonetary Incentives Would Attract Star Teachers?

Teacher surveys and focus groups consistently identified other aspects of teaching, besides salary, that influence teachers’ career decisions (Shapiro and Lane 2005; Ingersoll 2003). Many educators and policy experts we spoke with agreed that the star teacher program would need to motivate participants using more than bonuses. The idea of serving the most disadvantaged students might appeal to candidates’ sense of altruism, but the demonstration should seek additional ways to increase the attractiveness of the program by addressing any factors that may influence a teacher’s mobility decision.

We recommend making nonmonetary incentives, including professional development or school leadership roles, an integral part of the star teacher award. This section presents several types of nonmonetary incentives that an evaluator can consider using in the demonstration. We propose obtaining additional input from a panel of experts to refine the nonmonetary incentives.

The evaluator can seek ways to raise the prestige of participation in the program. Being singled out for distinction as a high-quality educator can be helpful in advancing a teacher’s career as a future administrator or school leader. For example, the Highly Skilled Educator Program in Kentucky successfully publicizes the high quality of its participants, which leads principals to recruit them based on the strength of their achievements. Similarly, other counties in Alabama have recruited teachers hired for Mobile County’s Transformation Schools program because “word has gotten around that they are good people to hire”
II: Star Teacher Program Design

(Rasberry et al. 2006). The program design should be careful using formal recognition because it has the potential of creating animosity among existing teachers in high-need schools who are categorically ineligible for the program but might otherwise have qualified.

Another potential incentive is the sense of service the program offers some teachers. The New Teacher Project successfully attracted 500 certified teachers in Oakland and 4,300 in New York City by using this social responsibility message to market the program (no financial incentive was offered).³ When asked why they applied for the Oakland program, applicants most frequently responded that the program’s “socially meaningful work” and its attempt to address inequality motivated their participation. Importantly, the New Teacher Project combined the message of meaningful work with a guarantee to place teachers in schools through an expedited hiring process--another intangible benefit of the program.

Opportunities for professional development can also make the demonstration more attractive to teachers. Professional development could range from a multiday training conducted by the star teacher program as preparation for teaching in high-need schools to stipends to fund more formal professional development opportunities. For example, Virginia requires each participating teacher to attend professional development and offers a $500 stipend to support such activity. Kentucky developed an innovative approach for its Highly Skilled Educator program in which participation in the program counted as credit toward certification as an instructional leader, principal, or superintendent.

The demonstration should consider ways to address teachers’ concerns about school working conditions. Teachers tend to cite a common set of working conditions that affect their mobility decisions, including a strong and committed principal, influence in school decision making, and support for student discipline (Shapiro and Lane 2005; Ingersoll 2003; Loeb et al. 2003). Experts we interviewed, including superintendents, repeated that a strong principal was pivotal to teachers’ willingness to work in a new school. We provide two potential approaches to addressing working conditions that depend on how the evaluation is designed. Each approach expands the scope of the star teacher program somewhat beyond a teacher transfer program to incorporate some changes that affect the whole school.

In the first approach, the demonstration could target schools in restructuring status that are experiencing wholesale change, including the rehiring of principals and staff and an infusion of additional training and resources. For example, Mobile County hired new principals for all five schools, required all staff to reapply for teaching positions, and offered selected schools additional professional development and other resources. Hamilton County, Tennessee, which encompasses Chattanooga, used a similar approach in reconstituting schools, beginning with principals.

³ The two New Teacher Project programs are the Oakland Teacher Corps and New York Urban Teachers (formerly called Excelsior Teaching Initiative). Information on these programs comes from interviews with two executive-level staff at the New Teacher Project.
In the second approach, teachers interested in having a broader decision-making role could take on an additional position in the school as an instructional leader, mentor, or school team leader. Several states have implemented programs that place teachers as instructional coaches rather than assigning them to teach in a classroom. For the demonstration, we propose somewhat scaled-down leadership opportunities that would provide teachers with the ability to influence the school at a higher level. Such an approach requires a between-school evaluation design, with entire schools assigned to a treatment or control condition (see Chapter III); otherwise, star teachers’ interaction with control teachers in the same school would mask the program’s impact.

Another aspect of working conditions relates to release time from teaching. Providing release time for additional leadership responsibilities in the school would enhance the attractiveness of star teacher positions but would be costly. However, star teachers could use release time for leadership roles in the school as discussed above, participation in professional development, or leadership of professional development or training. One expert we interviewed said that, in districts located near teacher training institutions, the program could arrange part-time adjunct faculty positions for star teachers. Such positions would mean increased teacher prestige and provide valuable career experiences, thereby strengthening the incentive to participate and producing additional benefits for the community.

3. Should Star Teachers Be Required to Make More Than a One-Year Commitment?

The star teacher program faces a tradeoff when considering the required duration of a participants’ commitment to a high-need school. Longer commitments yield greater potential benefits to the high-need school but generate greater program costs and increase the risk of low participation by star teachers. Furthermore, the demonstration itself will be more useful to the policy community if the evaluation is completed in a timely fashion. Ideally, the demonstration will measure the retention of star teachers in high-need schools and students’ learning gains at least one year beyond the recruitment year. In considering the most realistic options, we face a choice of requiring a two-year commitment from star teachers, a one-year commitment with an option for a second year, or simply a one-year commitment. In this section we explain our recommendation for a compromise position.

We recommend requiring star teachers to sign a one-year commitment while offering an opportunity to earn a bonus again in the second year.

Our main concern with a required multiyear commitment is the willingness of star teachers to commit to more than one year in a high-need school. A two-year commitment raises the cost of the program for prospective participants because of greater uncertainty about teachers’ potential return to their sending school and less flexibility if teachers are dissatisfied with their placement. The evaluator can reduce the costs of a two-year commitment by providing a strong guarantee that teachers will be able to return to their sending schools, but such a guarantee is unrealistic unless an agreement is negotiated with the teachers’ union. Alternatively, requiring a one-year commitment but offering a second
year of the bonus provides an incentive for star teachers’ return without the reduction in applicants that might result from a two-year requirement. The experience of Mobile County suggests that retaining star teachers in high-need schools may be a challenge, as 55 teachers left the program in its first two years despite their signing a five-year contract.

C. PLACING STAR TEACHERS IN HIGH-NEED SCHOOLS

The star teacher demonstration facilitates the transfer of highly effective teachers into high-need schools. Key elements of this process include establishing criteria to select high-need schools and designing a process to ensure the timely placement of star teachers. We provide recommendations for both aspects of the star teacher demonstration below.

1. How Should the Demonstration Define High-Need Schools?

The goal of the star teacher program is to raise student achievement by filling school vacancies with effective teachers. Therefore, the program is most relevant for schools with low student achievement that experience difficulty in attracting or retaining effective teachers. To ensure that the demonstration is appropriately targeted, the definition of high-need schools should address both school performance and quality of the teaching force. Since measures of teacher quality may not be readily available, proxy measures that indirectly address teacher quality should be used instead. Importantly, the definition of high-need schools should establish a large enough pool of eligible districts to make the evaluation feasible.

We recommend defining high-need schools based on schools that have not achieved AYP in some set of recent years (to be determined), have a high proportion of low-income students, and, if feasible, have high teacher turnover.

The definition of high-need school should include an indicator of school performance for target low-performing schools. States monitor school performance in terms of adequate yearly progress (AYP) toward the ultimate goal of universal proficiency by 2014. The use of AYP to define high-need schools would be particularly relevant for both state and district officials held accountable for raising achievement levels in schools not achieving AYP. The evaluator could use a measure based on the number of recent years a school has not achieved AYP or rely on the labels assigned to schools that have not achieved AYP over consecutive years, including “in need of improvement” after two consecutive years, “corrective action” after four years, and “restructuring” after six years. Each label is tied to a series of action steps and reforms that schools must implement to address low performance. In some cases, participating schools might see the intervention offered by the investigators carrying out the demonstration as a boon; the demonstration can help fulfill a requirement for taking corrective action and therefore would simplify the normally burdensome task of recruiting schools for the study.

Applying the AYP criteria would yield a promising number of districts with schools that meet the high-need criteria. Based on the most recent data from the 2004–2005 school year, just under 3,000 districts had a school identified for improvement (i.e., a school labeled as in
need of improvement, in corrective action, or in restructuring) (Stullich et al. 2006). While most districts had only one or two identified schools (73 percent), just under 800 districts had at least two identified schools, and 129 districts had 13 or more identified schools. Overall, the nation accounted for about 11,000 identified schools, including 9,000 Title I schools. The school performance criteria will overlap with the school poverty measure because schools with high proportions of low-income students were much more likely to be identified for improvement (Stullich et al. 2006). In addition, schools in urban areas and large school districts are much more likely to be identified.

One challenge of using AYP to define high-need schools is AYP’s lack of comparability across states. Though not a problem for implementing the demonstration, it is an issue to consider for the generalizability of the evaluation. State accountability systems differ in their methods of calculating AYP, their timing of AYP proficiency goals, the cut-off scores used to define proficiency, and the tests used to measure student achievement. Consequently, AYP measures do not translate across states, and differences in state accountability systems will account for some of the difference between high-need schools across states.

As mentioned, the high-need school criteria should also identify schools that have difficulty attracting and retaining effective teachers. While identifying the ability of schools to hire and keep effective teachers is not feasible, several other measures could serve as proxies.

**Teacher turnover rate.** The demonstration could target schools with historically high turnover, an indication that a school has difficulty retaining teachers. Combined with the school performance criteria, the study would focus on low-performing schools that struggle to retain teachers. This approach has the advantage of identifying schools that are likely to have the vacancies needed for the demonstration. The measurement of teacher turnover is dependent on adequate district data on the number of teachers exiting and entering each school; such data may be difficult to obtain. We propose teacher turnover as an additional criterion for defining high-need schools if the necessary data are available.

**Proportion of underqualified teachers.** Alternatively, the evaluator could examine the qualifications of a school’s teaching force to identify schools with a large proportion of unqualified or underqualified teachers. A school that has trouble recruiting and retaining effective teachers would have a teaching force disproportionately composed of uncertified teachers or teachers lacking content-area mastery. Teacher qualifications such as preparation, certification, experience, and degrees are perhaps weak indicators of teacher effectiveness in terms of ability to produce student learning gains, but they are somewhat easily measured criteria that feed into state and federal guidelines for “highly qualified teacher” status. High rates of out-of-field teaching or uncertified substitute teaching in regular classrooms would indicate a school’s high need for better-qualified teacher candidates.

**School poverty.** Given that high-poverty schools have high turnover rates and a disproportionate share of unqualified teachers, school poverty can serve as an indicator of schools that struggle to recruit and retain effective teachers. The research indicates that high-poverty schools have higher turnover rates as compared with other public schools.
(Ingersoll 2003; Hanushek et al. 2001; Shen 1997; Provasnik 2005). High-poverty schools are also more likely to report difficulty in filling vacancies and, as a result, rely more heavily on unqualified teachers to fill vacant positions (NCTAF 1997). Studies also find that high-poverty schools have a disproportionate number of teachers with lower qualifications in terms of factors such as teaching experience, certification, and master’s degree (Lankford et al. 2002; DeAngelis et al. 2005; Peske and Haycock 2006). For several reasons, a high poverty rate in a school can be an indicator of unfavorable working conditions. It may be correlated with high crime in the school area and with higher rates of learning or discipline problems.

The most commonly used measure of school poverty is the proportion of students receiving free or reduced-price lunch. Data on this measure are easy to obtain and have additional policy relevance because the allocation of Title I funds, the largest source of federal funding for schools, is based on these data. Neighborhood poverty is similar to school poverty and would indicate areas that are likely to have fewer local resources to hire and attract high-quality teachers. This measure could be developed with Census Bureau data.

A preliminary analysis of school poverty data from the Common Core of Data suggests an encouragingly large number of districts that may meet the high-need definition. We identified the number of districts with six or more schools with at least 50 percent low-income students (see Table II.3). For elementary schools, we counted 858 districts; for middle schools, 290 districts. The districts were fairly large, and at least half of the districts in the middle school analysis had 11 or more middle schools. At least half of the districts in the elementary school analysis had 13 elementary schools.

| Table II.3. Numbers of schools, districts, and states potentially available for study |
| --- | --- | --- |
| | Elementary Schools with at Least 50 Percent Low-Income Students | Middle Schools with at Least 50 Percent Low-Income Students |
| Eligible Schools | 19,481 | 5,217 |
| Districts with more than 5 eligible schools | 858 | 290 |
| States with at least one district having 5 eligible schools | 46 | 42 |
| Median number of eligible schools in districts with more than 5 eligible schools | 13 | 11 |
2. **What Are the Key Aspects of the Star Teacher Placement Process?**

The role of the star teacher program is to facilitate the transfer of highly effective teachers into high-need schools. The program essentially reverses a process that often works against the interests of high-need schools; teachers typically transfer out of high-need schools, but few teachers choose to transfer in to fill vacancies. This movement of teachers across schools within districts is typically a function of collective bargaining agreements with unions or a “meet and confer” process between teachers and the district. The evaluator can work within the existing teacher transfer process or negotiate with teachers to establish a separate placement process for star teachers.

Pending additional consideration from a panel of experts, we recommend using the voluntary teacher transfer process to match teachers with high-need schools, ensuring that star teachers are placed by early May. A recent study explains how the transfer process can delay the hiring process and place urban districts at a distinct disadvantage for hiring new teachers and other teachers from outside the district (Levin and Quinn 2003). However, the process works to the benefit of star teachers transferring within the district because they could interview for vacancies in high-need schools before the positions are opened to outside candidates. In addition, if star teachers are more experienced than other teachers, the voluntary transfer process may give them priority in being placed in a position if union rules give preference based on seniority.

The structure and timing of the teacher transfer process can vary widely across districts. Therefore, we present important aspects of the transfer process that the evaluator should consider when determining the feasibility of using the teacher transfer process to facilitate the placement of star teachers.

**Identifying vacancies.** The evaluator should identify vacancies in high-need schools early enough to ensure that star teachers can take advantage of the teacher transfer process. Early identification may be a challenge because principals in large urban districts sometimes hide their vacancies to avoid filling the position with a teacher who lost their current teaching placement due to enrollment or budget declines (Levin et al. 2005). Retiring teachers may not report their retirement plans to the district because of disincentives associated with early notification. Therefore, the evaluator should coordinate efforts with both the district human resources department and principals in high-need schools. Maintaining close communication with principals can assist in identifying vacancies. The opportunity to fill a vacancy with a star teacher may encourage principals to share information about potential vacancies.

**Principal discretion in hiring voluntary transfers.** The use of the teacher transfer process to place star teachers in high-need schools assumes that principals have the ability to select between teachers interested in the position. However, union rules may impose restrictions on which voluntary transfers a principal can choose from or which transfer the...
principal must hire. For example, principals in five urban districts reported that they have no choice or restricted choice in selecting teachers for 40 percent of all vacancies (Levin et al. 2005). A majority of schools in each of the study districts filled at least one vacancy without any choice or with restricted choice. Union rules may require principals to hire from a seniority-based list of vacancies or, in some cases, to hire the teacher with the greatest seniority. The seniority issue is not a problem if high-need schools have few teachers interested in filling the vacancy, but the evaluator should consider issues related to seniority when planning the placement process.

**School location.** Star teachers may be more willing to accept teaching positions that are nearby and do not require relocation or a lengthy commute. A recent study found that geographic location is a factor in teacher mobility decisions (Boyd et al. 2005). Therefore, the evaluator should consider proximity when matching star candidates to participating high-need schools for interviews.

**Timely placement.** The timing of star teacher placement into high-need schools will affect teachers’ willingness to participate in the program. A study of five urban districts studied by the New Teacher Project found that high-quality teachers are unwilling to wait until late in the summer to complete the hiring process (Levin et al. 2005). For example, the five districts in the study lost 30 to 60 percent of their new teacher candidates, and the large majority of these applicants cited the late hiring timeline as the reason they took a position elsewhere. The study suggests that urban districts need to complete the hiring process by May in order to attract and hire high-quality candidates. While the New Teacher Project study focused on teachers who applied from outside the district, we suspect that highly effective teachers within the district will also respond unfavorably to a late hiring process.

The proposed timeline (Table II.4) is structured to reflect a transfer process that begins by April and concludes by early May. It is entirely feasible because the teacher transfer process occurs early enough in the school year. As an example, all but one of the five urban districts included in the New Teacher Project study began their transfer process before May. This timing affords the demonstration enough time to prescreen teachers, conduct recruitment, identify vacancies, and select star teacher finalists by April, when interviews should begin. The one contingency is the identification of vacancies. If principals are not aware of their vacancies by then or if vacancies are not determined until later in the school year, lack of information could delay the start of interviews.

<table>
<thead>
<tr>
<th>December</th>
<th>January</th>
<th>February</th>
<th>March</th>
<th>April</th>
<th>May</th>
<th>June</th>
<th>July</th>
<th>August</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prescreen teachers using value-added analysis</td>
<td>Publicize program and recruit teachers</td>
<td>Identify school vacancies and conduct star teacher selection process</td>
<td>Begin teacher interviews and placement</td>
<td>Complete teacher interviews and placement</td>
<td>Recognize selected teachers</td>
<td>Orientation / training</td>
<td>Star teachers begin teaching</td>
<td></td>
</tr>
</tbody>
</table>
3. **Should Star Teachers Be Placed Into High-Need Schools Individually or as a Group?**

Placing groups of star teachers, rather than individual teachers, in high-need schools offers an opportunity to have a broader school-level impact while offering greater support for star teachers. Group placements may improve the willingness of teachers to participate because a cadre of highly effective teachers would substantially change the school working environment. The major challenge of placing groups of star teachers is finding enough schools that have multiple vacancies in the targeted grades. In addition, the decision between individual or group placements depends on whether the study conducts random assignment within or between schools.

*We recommend placing groups of star teachers in high-need school if the evaluation chooses a between-school design.*

There are several advantages of assigning groups of star teachers to the same school. Groups of star teachers would have an immediate support network within their school and fellow colleagues with whom they could collaborate and share resources. Teach for America relies on this approach, and last year placed over 90 percent of its teachers in schools that already had a current or former Teach for America teacher. There is also a greater potential for school-wide impacts if a group of highly effective teachers contributes to more sharing of knowledge between the star and incumbent teachers, or leads to a change in the school culture. Principals and teachers in Mobile County described a distinct change in the school environment of schools where large groups of participating teachers were hired through their bonus program (Rasberry et al. 2006).

**D. ADDRESSING KEY STAKEHOLDERS**

Implementation of the star teacher demonstration requires the buy-in of several key stakeholders, including teachers’ unions, the schools where star teachers are currently teaching (sending schools), the high-need schools that fill their vacancies with star teachers (receiving schools), and host districts. This section discusses the perspective of each stakeholder group and its potential concerns.

1. **What Are the Potential Concerns of Teachers’ Unions?**

Experience suggests that implementing teacher pay initiatives without the input or support of teachers can lead to conflict with teachers’ unions or opposition from teachers. Indeed, case studies of several forms of teacher compensation reform found that developing a cooperative relationship with teachers’ associations was a key factor in successful implementation of reform initiatives (CPRE 2006). The country’s two largest unions may be receptive to the concept of the star teacher program because they have both expressed support for incentives to attract teachers to hard-to-staff schools (AFT 2002; Weaver 2006). However, those same unions strongly reject reforms that differentiate teacher pay based on performance and are likely to express concerns about defining eligibility for the program based on teacher effectiveness.
We recommend engaging teachers’ unions early in the process of implementing the star teacher demonstration to identify and address teacher concerns.

A primary concern of teachers’ unions is that financial incentives represent only one component of the broad change required to raise achievement in high-need schools. Unions argue that, in addition to attracting qualified teachers, incentive programs should address working conditions within high-need schools, including strong and supportive leadership, security, and professional development (AFT 2002; Delisio 2006). Unions are concerned that teachers cannot turn around high-need schools by themselves and should not be held responsible for doing so; turnaround requires broader change. Three of the four programs we identified that resemble the star teacher program combined their incentives with additional interventions in the school. While professional development was the most common intervention, Mobile County included its incentive as part of a broader reconstitution effort. The five participating schools rehired for all positions in the schools while the district provided instructional coaches, school-wide professional development, and additional support staff.

Teachers’ unions have made it clear that they do not approve of performance-based pay. While the star teacher demonstration is an incentive for high-need schools, eligibility for the program is based on performance. That aspect of the program can give rise to concerns about equity and lead to the argument that all qualified teachers should be eligible for the program. Given that the star teacher bonus is not based on performance (i.e., star teachers receive the bonus without regard to their performance), such an incentive may not generate much concern.

Teacher compensation and the teacher transfer process are typically covered by collective bargaining agreements between districts and unions (or other teacher associations). Since the star teacher incentive represents a change in compensation for teachers, the evaluator may need to seek the approval of the union. Such was the case in Palm Beach County, where the district signed a Memorandum of Understanding with the teachers’ union to obtain approval for the $10,000 bonus awarded to eligible teachers who transferred into low-performing schools. In addition, if the evaluator establishes an expedited teacher transfer process or plans to revise or work outside the process, the union may have to grant approval.

District administrators may have concerns about the star teacher program. We interviewed one expert who had designed and implemented several teacher pay reforms in an urban school district. He warned that the strongest resistance to change might come not from teachers’ associations but from instructional leaders at the district level, such as the superintendent of instruction or chief academic officer. He noted that such resistance is often rooted in ideological or philosophic objections to programs that resemble merit pay or differentiated pay for any types of teachers with equal experience and education.

The other objections of school district officials would be more practical. District officials weigh the tradeoff between the added resources and added burdens imposed by the study. The added resources would come in the form of teacher bonuses that may boost the attractiveness of teaching in that district. Such an inducement would be stronger, of course,
II: Star Teacher Program Design

if the demonstration followed the between-district program design, which would use resources to attract talented teachers from other districts instead of shifting them around within the district. One type of cost to districts might be borne through compliance with random assignment, i.e., excluding some schools from the study’s supposed benefits. If the study randomly assigns students within schools, this pushes the burden of complying with random assignment down to the school. Another cost to districts is associated with facilitating the intervention (using staff hours from the human resources department to accommodate teacher transfers and other matters) and participating in data collection (requiring staff hours from research and evaluation offices). These burdens can sometimes be offset with incentive payments commensurate with the effort, but they need to be considered a cost of doing business with overtaxed school districts.

2. How Will Sending Schools Respond to the Program?

The demonstration relies on low-need schools for its supply of highly effective teachers. The justification for encouraging star teachers to transfer from these sending schools is twofold: (1) low-need schools have a disproportionate share of effective teachers, and (2) high-need schools have a more difficult time attracting these teachers. Despite these assumptions underlying the star teacher program, sending schools will resist the loss of a teacher, especially after the program identifies the teacher as highly effective. While the sending school cannot prevent a teacher from transferring, the principal could offer incentives to induce the teacher to remain (e.g., a preferred grade or subject placement or a school leadership role). Additionally, the evaluator may need the voluntary cooperation of sending schools to communicate with and access star teachers.

We recommend exploring the possibility of offering the sending schools a nonmonetary incentive, such as preferred treatment in the teacher transfer or hiring process to refill their vacancy.

Two aspects of the star teacher program are likely to affect sending schools’ opposition to the demonstration: the length of the teaching commitment and the number of star teachers leaving the sending school. A longer commitment may cause sending schools to resist the program if they expect teachers to return after they complete the program. Moreover, if star teachers are concentrated in a small number of schools, the loss of several teachers from a school could be disconcerting. The evaluator should be prepared to address the potential concerns of schools with a relatively large number of star teachers. Recruiting large, diverse districts for the study could provide a broader distribution of star teachers across schools.

Sending schools may be motivated to participate in the program if the evaluator and/or district makes a strong case for why the high-need schools need the teachers. For example, in Palm Beach County, a labor relations official suggested that principals from sending schools were willing to send their highly effective teachers because they supported the program’s goal of raising achievement in the district’s lowest-ranked schools (referred to as “F schools”). The stigma attached with F schools generated a greater willingness to allow the transfer of participating teachers. The evaluator should clearly communicate the purpose
and goals of the program to sending schools and effectively define the need for stronger teachers in low-need schools.

3. **What Are the Potential Concerns of Receiving Schools?**

Star teachers will be placed in low-performing and high-poverty schools that struggle to retain effective teachers. While school administrators will probably be eager to hire star teachers for the demonstration, teachers in the receiving school may be less welcoming. Incumbent teachers may view differential pay for teachers in the same school as unfair, especially as the existing teachers were ineligible to apply for the award. The experience of Palm Beach County suggests that, even in schools where teachers are eligible to receive the award, incumbent teachers may express resentment. Teachers in the receiving schools may feel that the placement of a star teacher in the school questions their own commitment or teaching abilities while not adequately recognizing the challenging environment of a high-need school. Resentment between incumbent and star teachers could create barriers that isolate star teachers in their classroom and undercut teacher collaboration and coordination—which does not create a productive learning environment in the school.

*Therefore, we recommend a short orientation or training period for star teachers to prepare them for their placement in a high-need school.*

Addressing the potential for conflict between incumbent and star teachers is important. Such conflict can affect the impact of star teachers on their classrooms and schools as well as teachers’ willingness to stay in the program for a second year. A training or orientation can provide guidance for star teachers on ways to integrate into the challenging environment of a high-need school. A discussion facilitated by the Center for Teacher Quality with National Board–certified teachers placed in high-need schools recommended that teachers adopt a “learning stance” by making efforts to learn about the school, its students, and the community (Berry and King 2005). These teachers warned that trying to distinguish themselves as experts may result in isolation and generate resistance from existing teachers. Such strategies can encourage positive relationships between incumbent and star teachers. The New Teacher Project conducted similar training for certified teachers placed in high-need schools in its Oakland and New York City programs.

---

5 Earlier in this chapter, we recommended that teachers in low-performing schools should be ineligible for the program. Since low performance is also a criterion for defining a high-need school, teachers in the receiving school should be ineligible for the program.
CHAPTER III

EVALUATION OF THE STAR TEACHER PROGRAM

This chapter identifies design issues that must be addressed in a rigorous evaluation of the star teacher program. The discussion of these issues is structured around four central questions:

- What is the most policy-relevant counterfactual?
- What should be the unit of random assignment?
- How many schools will the evaluation require?
- What data will the evaluation require?

The answers to the questions are interdependent and they hinge on answers to questions raised in the program design discussion (Chapter II). For example, the method for approximating the counterfactual might differ according to whether it is feasible to randomly assign either whole schools or students within schools, and the sample size requirements depend, in turn, on the unit of random assignment and program design decisions that were discussed in Chapter II. We discuss these questions one at a time for the sake of simplicity.

A. THE COUNTERFactual

Any rigorous impact evaluation is intended to provide an estimate of the effects of a policy change by comparing outcomes realized under that change to what would have occurred absent that change or, in other words, had the status quo prevailed. That condition is the “counterfactual.” Since the counterfactual cannot be observed, the evaluation uses a control group to approximate the counterfactual state.

In the context of the star teacher evaluation, the most policy-relevant counterfactual and the appropriate control group for approximating that counterfactual are not immediately obvious. We explore the issues involved with the choice of counterfactual by discussing
each of two central questions in turn: (1) What is the most policy-relevant counterfactual?; and, (2) What is the most feasible approach to approximating the counterfactual? As explained below, we recommend choosing a counterfactual state that represents the natural mix of teachers encountered in high need schools in the absence of the star teacher program and approximating this counterfactual using a control group that includes both newly hired and returning teachers.

1. What Is the Most Policy-Relevant Counterfactual?

To the extent that all schools potentially interested in offering an incentive package to recruit star teachers have a vacancy that needs to be filled, the most policy-relevant counterfactual state is one in which schools would fill a vacancy by hiring a new teacher. An evaluation that chooses this counterfactual would seek to compare the performance of star teachers to the performance of the newly hired teachers who would have taught study students in the absence of the star teacher program.

Focusing solely on newly hired teachers, however, may not fully capture the situation that would prevail under the status quo. Instead of hiring a new teacher, the school could entice an existing teacher who is at risk of leaving the school to return for another school year. Comparing the star teachers’ performance to that of returning teachers would thus be of interest if one assumes a counterfactual in which study students would have been taught by returning teachers in the absence of the star teacher program.

Since schools have the ability to exercise both mechanisms for filling a vacancy—hiring a new teacher or retaining an existing teacher—the state that would prevail in the absence of the star teacher program would arguably contain a mix of new and returning teachers. Choosing a counterfactual that captures this natural mix could provide a more complete picture of the effects of the star teacher program.

2. What Is the Most Feasible Approach to Approximating the Counterfactual?

One approach to approximating a counterfactual state in which study students would have been taught by new hires in the absence of the program is to use a control group consisting solely of newly hired teachers. The outcomes of star teachers would be compared to the outcomes of other teachers working in the same grade and subject who are new to high need schools. The main difference between treatment and control teachers in this case would be their years of experience in the profession. Teaching experience would be confounded with treatment effects, a confounding that may be desirable, depending on how one wishes to interpret the impacts.

In terms of feasibility, however, designing the evaluation around new hires raises some concern. The participating schools would have to have more vacancies to be eligible for the study because there must be enough slots to accommodate the treatment as well as the control teachers. While the target schools are expected to have high turnover, this may still disqualify many schools from the study and make school recruiting more difficult. (See below for further discussion of this issue.)
An alternative approach that would have greater chance of success while also allowing for the comparison of star teachers to returning teachers would be to use a control group that includes a combination of newly hired and returning teachers. Chances of success would be greater because the less stringent requirements for the comparison to returning teachers enlarges the pool of schools eligible to participate. Using a control group that combines new and returning teachers creates an opportunity for estimating the impact under each comparison separately, with the total impact describing the combination of both types of teachers (new and returning) that students would face.

This mixed control group would include teachers already working in the same grade and subject at high-need schools. One main difference between treatment teachers and returning control teachers would be their tenure in the high-need schools; star teachers presumably would have zero years of experience in study schools, whereas returning teachers would have at least one year of experience in study schools. School seniority would be confounded with treatment effects, but that may be desirable, depending on how one wishes to interpret the impact estimates. For example, if we were to compare treatment teachers (who by construction have at least four years’ experience) to control teachers with no experience, then a positive program impact would be dismissed by detractors of the program approach as a simple experience effect and not an effect of the successful identification or recruitment (with incentives) of talented “star” teachers.

As a preliminary recommendation, we suggest targeting a combination of newly hired and existing control teachers in high-need schools. We believe both comparisons are relevant to policy, as the variation in teacher quality across schools appears to apply across a range of teaching experience and school tenure levels as well. Stated somewhat differently, students in high-need schools could suffer the consequences of a low-quality education regardless of whether they are taught by a newly hired or a returning teacher, and the study would be more useful if it could address both possibilities.

3. Implications for Evaluation Design

The evaluation team will need to consider the preceding confounding effects in interpreting the results. Ideally, the evaluator would recruit schools such that control group teachers represent a range of overall teaching experience and school tenure levels. The number of years of teaching experience and the number of years at the current school could be controlled for. Using nonexperimental methods, the evaluation team could then assess whether and how the estimated impacts of the star teacher program vary with the experience and school tenure of the control group teachers. The ability to identify fine variations in the treatment effect will be constrained, however, by both the sample size and the extent of variability in the data.

B. Unit of Random Assignment

Random assignment can be imposed at a variety of levels. In this section, we discuss the advantages and disadvantages associated with random assignment at the student level.
and at the school level. Several of the issues discussed here will be revisited in Section C, where we quantify aspects of the tradeoffs between the two designs.

1. Randomly Assign Students

One design option is to randomly assign students to either a star teacher or a control teacher in the same grade. The classroom of students taught by the star teacher would constitute the treatment group, and the classroom of students taught by the control teacher would constitute the control group. Comparing the outcomes of the students in the two classrooms would yield unbiased estimates of the impact of the star teacher program on student achievement. Comparing the qualifications of the two teachers and the number of years they remain in the school after the study begins would yield unbiased estimates of the impact of the program on recruitment and retention, respectively.

The main advantage of random assignment of students is that a comparison of classrooms within the same school rules out school-level factors as an influence on outcomes. In the context of the star teacher evaluation, holding constant the characteristics of the principal may be of particular importance. Evidence from the literature and our conversations with practitioners suggest that the principal may be a primary factor affecting the willingness of a star teacher to transfer to and remain in a high-need school. The annual survey of new teachers at schools participating in the Benwood Initiative, for example, revealed that teachers most frequently cited the principal as the factor most important in their decision to accept a position at a Benwood school (Handley and Kronley 2006). To the extent that principal characteristics vary across schools and influence student and teacher outcomes, holding the principal constant by using a within-school random assignment design would improve the precision of the estimated impacts of the star teacher program.

The main disadvantage of random assignment of students to teachers is the possibility of “contamination” or “spillover” effects. Contamination would occur if control group teachers change their behavior in response to the presence of the star teacher in the school or if control group students otherwise receive some benefit from the presence of star teachers. For example, control group teachers might implement teaching methods learned from collaborating with or observing star teachers. Control group students might be reassigned to star teacher classrooms. Such contamination would affect the counterfactual and bias the impact estimates.

The evaluation could reduce the risk of contamination by imposing restrictions on the interactions between star teachers and control teachers. Such restrictions, however, could hinder the ability of the program to attract star teachers; in fact, school professionals indicate that the opportunity to serve in a leadership role and engage with other teachers might weigh heavily in a star teacher’s decision to transfer to a high-need school. Furthermore, pressures on school staff to undermine the random assignment of students might still exist. Consequently, this design requires considerable monitoring of school activities in order to maintain the integrity of random assignment.

III: Evaluation of the Star Teacher Program
2. Randomly Assign Schools

An alternative design is to randomly assign schools to the treatment and control conditions. Schools randomly assigned to the treatment condition would be eligible to offer the star teacher incentive package, and schools assigned to the control condition would not be eligible to do so. The evaluation would compare outcomes of teachers and their classrooms across treatment and control schools to produce unbiased estimates of the impact of the star teacher program.

One major advantage of the random assignment of schools to treatment (a between-school design) is that the design greatly reduces the risk of contamination without imposing restrictions on teacher collaboration, perhaps making the program more palatable to schools and teachers. While students could be reassigned to classrooms within schools, they are far less likely to transfer to another school in order to be placed in a star teacher classroom, thus further reducing the possibility of contamination bias.

Another relative advantage of random assignment at the school level is that it makes school recruitment easier. Unlike random assignment at the student level, random assignment of schools does not require permission to interfere in the student-teacher matching process, likely making principals more willing to cooperate. School recruitment in the between-school design is also easier because, depending on the desired counterfactual comparison, it could require fewer vacancies per school. Table III.1 shows the minimum number of vacant teaching slots that a school (or school-grade combination) needs for each of several possible study designs if it is to participate in the study. If we wish to compare star teachers to returning teachers only (Designs IA and IB), then only one teaching vacancy is required regardless of the unit of random assignment. The unit of random assignment does matter, however, for the other types of comparisons. If the control group consists solely of newly hired teachers, a student-level random assignment design (Design IIA) requires at least two slots within the same subject and grade so that at least one slot is available for a star teacher and at least one slot is available for another new hire to whom the star teacher will be compared. School-level random assignment with the newly hired control group (Design IIB) still requires only one vacancy in each school; that vacancy can be filled by a star teacher for a treatment school or a newly hired teacher for a control school. The requirements for the combination control group (Designs IIIA and IIIB) must be the same as those for newly hired teachers for at least a subset of schools in order to permit a comparison to new hires. Thus, school-level random assignment can accommodate all three control groups with only one vacancy in each school.

Random assignment of schools offers several other advantages. The aforementioned advantages of reduced risk of contamination and the requirement of only one vacancy per school broaden the pool of schools eligible for the evaluation. For example, school-level random assignment can more easily accommodate schools with compartmentalized teaching or small schools with only one vacancy in each grade. Based on our experience in other studies, we believe that the recruitment of schools and monitoring of random assignment
Table III.1. Minimum Number of Teaching Slots per School-Grade Required by Design

<table>
<thead>
<tr>
<th>Units of Random Assignment</th>
<th>A. Students</th>
<th>B. Schools</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Returning Teachers</td>
<td>Design IA. 1 slot</td>
<td>Design IB. 1 slot</td>
</tr>
<tr>
<td>Control Group Used to Approximate the Counterfactual</td>
<td>II. Newly Hired Teachers</td>
<td>Design IIA. 2 slots</td>
</tr>
<tr>
<td>III. Combination</td>
<td>Design IIIA. 2 slots in at least a subset of schools</td>
<td>Design IIIB. 1 slot</td>
</tr>
</tbody>
</table>

will be easier with a school-level design; Section C discusses the lower recruitment and monitoring costs associated with this design. For all of these reasons, random assignment at the school level would also expedite an exploration of the impact of recruiting a group of star teachers.

The key disadvantage of randomly assigning schools is the loss of statistical precision in the impact estimates and the subsequent need to include a larger sample of schools relative to the within-school random assignment approach. Random differences between schools would make detection of program impacts more difficult. In order for the evaluation team to disentangle the effects of the star teacher program from the influence of school-level characteristics, more schools would need to be recruited, thus increasing the cost of the evaluation.

The loss of precision and increase in the number of required schools could be mitigated to some degree by using a randomized block design, which is analogous to stratification techniques used to make statistical sampling more efficient. Under block random assignment, pairs or groups of schools would first be matched on “blocking factors” based on school characteristics. Schools within each matched pair or group would then be randomly assigned to the treatment or control condition. For example, school district could serve as a primary blocking factor so that random assignment of schools within districts would hold constant district policies that could affect outcomes. Other measurable school characteristics considered crucial to the outcomes—possibly including principals’ characteristics—could also be used as blocking factors.

3. Seek More Expert Advice Before Committing to a Design

Ultimately, the selection of students or schools as the unit of random assignment depends on the evaluation of the risk of contamination bias associated with the random assignment of students to teachers within schools, the challenges involved with recruiting schools for a student-level random assignment evaluation, and the costs of randomly
assigning schools. The power analysis in the next section concludes with an illustration of cost differences between the designs and thus helps inform the critical evaluation decision about the unit of random assignment. We anticipate that the expert panel will provide useful insights into the key factors that should be weighed in choosing a design, including the likelihood of spillover effects, the consequences of restricting teacher interactions to prevent contamination, and the importance and feasibility of controlling for principal characteristics.

C. SAMPLE SIZE AND POWER

As noted in the previous section, a school-level random assignment design offers several advantages over a student-level design, but at the cost of a precision loss that increases the required sample size. This section quantifies the precision tradeoff to provide a framework that can inform the evaluation design decision. We first estimate the number of schools required to obtain a given precision standard under each design. We then explore the differences in design costs and present an illustrative example showing how cost differences can affect the precision tradeoff. The illustration suggests that the lower costs associated with several aspects of the school-level design substantially mitigate the additional costs imposed by the larger sample size relative to the student-level design.

1. How Many Schools Are Required?

Samples sizes required to achieve a specified precision standard can be estimated by applying conventional formulas to a set of assumptions about the structure of the evaluation and the variability of the outcomes of interest. We present estimated sample sizes for both designs under consideration: random assignment of students to teachers within schools (Design A) and random assignment of schools (Design B). It is important to note that both designs involve cluster random assignment. Design B clearly involves clusters of schools. Though students are randomly assigned to teachers in Design A, the design involves clusters of classrooms because the treatment occurs at the classroom level—that is, an entire classroom of students will be either a treatment classroom taught by a star teacher or a control classroom taught by a control teacher. Consequently, the differences in effective sample sizes are not as stark as they might appear, where effective sample size is determined by the number and size of the clusters as well as the variability between versus within clusters.

The required sample sizes depend on the precision standard adopted for estimating the impact on student achievement. The precision standard may be stated in terms of a target minimum detectable effect (MDE), which represents the smallest impact, measured in effect size units (that is, as a percentage of the standard deviation of the outcome) that can be detected with high probability. Adopting a higher precision standard means that the
evaluation will be designed to detect a smaller impact, but at the cost of requiring a larger sample size. Table III.2 presents sample sizes for target MDEs ranging from 0.10 to 0.25.6

Table III.2. Sample Size Requirements by Design

<table>
<thead>
<tr>
<th>Design</th>
<th>MDE = 0.10</th>
<th>MDE = 0.15</th>
<th>MDE = 0.20</th>
<th>MDE = 0.25</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: Randomly Assign Students</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICC = 0.10</td>
<td>109</td>
<td>48</td>
<td>27</td>
<td>17</td>
</tr>
<tr>
<td>ICC = 0.15</td>
<td>142</td>
<td>63</td>
<td>36</td>
<td>23</td>
</tr>
<tr>
<td>ICC = 0.20</td>
<td>175</td>
<td>78</td>
<td>44</td>
<td>28</td>
</tr>
<tr>
<td>B: Randomly Assign Schools</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICC = 0.10</td>
<td>359</td>
<td>160</td>
<td>90</td>
<td>57</td>
</tr>
<tr>
<td>ICC = 0.15</td>
<td>496</td>
<td>220</td>
<td>124</td>
<td>79</td>
</tr>
<tr>
<td>ICC = 0.20</td>
<td>633</td>
<td>281</td>
<td>158</td>
<td>101</td>
</tr>
</tbody>
</table>

Note: We report sample sizes in terms of schools for the sake of simplicity. As discussed in the text, including multiple grades per school could reduce the number of schools required.

Assumptions

Number of teachers per school is 2 for Design A and 1 for Design B.

50% of within-school variance in test scores is explained by covariates (including pretest).
10% of between-school variance in test scores is explained by covariates.
0.50 correlation between treatment and control classroom means within the same school (applies to Design A only).
23 students per classroom.
Baseline and follow-up data are both available for 80% of students.
Two-tailed hypothesis tests conducted with 80 percent power and $\alpha = .05$.
Balanced sample design (50/50 split between treatment and control).

6 To obtain guidance on a precision standard for student achievement that is both meaningful to policymakers and realistically attainable within the constraints of the evaluation, we reviewed previous research and practices. Based on the results of meta-analyses conducted across a range of disciplines (for example, Cohen 1988; Lipsey and Wilson 1993), a common practice is to adopt an MDE of 0.20 as a precision standard. However, some evidence suggests that a smaller effect size may be both meaningful and attainable when estimating impacts on student achievement. Evidence from studies of elementary grades, for example, suggests that an effect size of 0.10 would indicate that a star teacher recruited to a high-need school increased the academic growth of his or her students by about 1 to 1.5 months, on average, relative to the counterfactual (Schochet 2005). Nonexperimental evidence from analyses conducted by Hanushek et al. (2005) suggested that a one standard deviation increase in teacher quality—which is the equivalent of comparing the achievement of a student whose teacher ranks at the 50th percentile in the quality distribution to the achievement of a student whose teacher ranks at the 85th percentile in the quality distribution—raises student achievement by at least 0.22 standard deviations. Related work (Rivkin, Hanushek, and Kain 2005) indicated a lower bound of 0.11 standard deviations. Experimental evidence from a rigorous evaluation of Teach For America (TFA), a program that aims to improve student achievement by placing high-achieving recruits deemed to be potentially effective teachers in low-income communities, found the impact of the assignment to a TFA teacher on mathematics achievement to be 0.15 effect size units (Glazerman, Mayer, and Decker 2006). The expert panel can provide further guidance in setting an appropriate precision standard.
Under typical assumptions, achieving a precision standard within the specified range requires about 3.5 times as many schools under the school-level random assignment design as under the student-level random design. We allow the intraclass correlation (ICC)—that is, the between-school variance in the outcome as a proportion of the total variance in the outcome—to vary to reflect a degree of uncertainty about this parameter.\(^7\) For example, obtaining an MDE of 0.20 when the ICC is 0.15 requires 36 schools under Design A compared to 124 schools under Design B.

We note that the reported sample sizes reflect calculations under a basic design in which there are only two teachers (one control and one treatment) in each school under Design A and only one teacher in each school under Design B. The tool we have developed for comparing sample sizes can readily accommodate more complex design elements, such as a cadre of star teachers in each school.

While we have reported sample sizes in terms of schools for the sake of simplicity, we should emphasize that the number of schools required could be reduced by including multiple grades per school. A sample size unit can be thought of as a school grade combination, such as grade 8 in school X. Several school grade combinations could be housed in a single school building. Including several grades per school in the evaluation would provide opportunities for increasing precision and reducing recruitment and data collection costs. The potential gains in precision are likely to be particularly large for Design B, because multiple grades per school enables the evaluator to control for fixed school characteristics that could not be controlled for under school-level random assignment if only one grade per school were included in the evaluation. Thus, the increase in the number of schools required for Design B relative to Design A would be smaller than the increases indicated in Table III.2.

2. Differences in Costs

In quantifying the precision tradeoff between the designs, it is important to note that some aspects of design implementation are less costly when random assignment is conducted at the school level rather than at the student level. Specifically, recruitment and monitoring costs tend to be substantially lower for school-level random assignment. Random assignment at the student level requires schools to grant the evaluation team control of the assignment of students to classrooms. Many schools perceive such a requirement as intrusive and fear that working with the evaluators to carry out the design will be excessively burdensome. Given that school-level random assignment does not require evaluators to assume control over within-school processes, the recruitment effort associated with a school-level design tends to be far less intensive and expensive than the recruitment for a student-level design. The student-level design also necessitates more frequent monitoring activities to protect the integrity of the random assignment, such as regularly cross-checking class rosters for crossover of students between treatment and control.

\(^7\) An ICC of 0.15 aligns most closely with previous MPR studies. We also present estimates for a more optimistic ICC value of 0.10 and a more conservative value of 0.20.
classrooms. Since there is considerably less scope for student crossover or other sources of contamination across schools, monitoring costs are expected to be lower for a school-level random assignment design.

3. Cost-Precision Tradeoffs Illustrated

To provide a more complete sense of the precision tradeoff, we present an illustrative comparison of the student-level (Design A) and school-level (Design B) random assignment designs that incorporates differences in both sample size requirements and variable costs (see Table III.3). In the illustration, we calculate the estimated variable costs associated with the sample size requirements shown in Table III.2. We should emphasize that the cost figures are intended for illustrative purposes only; more research is required to obtain precise cost estimates. To capture the potential differences in costs between designs, we assume the following: recruitment costs per school are $15,000 for Design A and $5,000 for Design B; monitoring costs per school are $5,000 for Design A and $500 for Design B. We assume that all other costs are the same for both designs. We first consider as an example the sample size requirements shown in Table III.2 for a target MDE of 0.20 and an ICC of 0.15. Under these assumptions, the total variable design cost is $1,584,000 for Design A and $2,294,000 for Design B. Thus, though requiring more than three times the number of schools, achieving an MDE of 0.20 with a school-level random assignment design costs only 45 percent more than achieving the same precision standard with a student-level random assignment design in this illustrative scenario. If we perturb the target MDE and ICC assumptions over the ranges shown in Table III.2 to reflect reasonable alternatives, the cost ratio varies from a low of 1.38 to a high of 1.52.

D. DATA COLLECTION AND MEASUREMENT OF OUTCOMES

To complete the discussion of the evaluation design, we outline the types of data that will be required to answer the study’s main questions. The data include measurement of outcomes as well as process variables, program implementation variables, and baseline/background variables that provide context for interpreting program impacts, performing subgroup analysis, and increasing the precision of estimates.

---

8 The illustrative estimates of total design costs presented in Table III.3 represent estimates of the variable costs associated with each design. We do not include the fixed costs that would be associated with either design.

9 In particular, we assume for the purposes of illustration that data collection costs are $2,000 per school and $1,000 per teacher. We also assume that treatment teachers—that is, star teachers, whom we assume represent half of all teachers in each design—receive $20,000.
Table III.3. Illustration of the Cost-Precision Tradeoff

<table>
<thead>
<tr>
<th>Unit of Random Assignment (Design A: Students) (Design B: Schools)</th>
<th>ICC</th>
<th>Target MDE</th>
<th>Estimated Sample Sized Required a</th>
<th>Estimated Costs for Illustrative Example b</th>
<th>Cost Ratio Design B/Design A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Students</td>
<td>.10</td>
<td>.10</td>
<td>109</td>
<td>$4,796,000</td>
<td>1.38</td>
</tr>
<tr>
<td>Schools</td>
<td>.10</td>
<td>.10</td>
<td>359</td>
<td>$6,641,500</td>
<td></td>
</tr>
<tr>
<td>Students</td>
<td>.15</td>
<td>.20</td>
<td>36</td>
<td>$1,584,000</td>
<td>1.45</td>
</tr>
<tr>
<td>Schools</td>
<td>.15</td>
<td>.20</td>
<td>124</td>
<td>$2,294,000</td>
<td></td>
</tr>
<tr>
<td>Students</td>
<td>.20</td>
<td>.25</td>
<td>28</td>
<td>$1,232,000</td>
<td>1.52</td>
</tr>
<tr>
<td>Schools</td>
<td>.20</td>
<td>.25</td>
<td>101</td>
<td>$1,868,500</td>
<td></td>
</tr>
</tbody>
</table>

Notes:

a See Table III.2 notes for other assumptions underlying the sample size calculations.
b Estimated design costs are presented solely for illustration. The illustrative cost figures assume the following:
- Recruitment costs per school: $15,000 for Design A; $5,000 for Design B
- Monitoring costs per school: $5,000 for Design A; $500 for Design B
- Data collection costs per school: $2,000 for both designs
- Data collection costs per teacher: $1,000 for both designs
- Incentive offered for each star teacher: $20,000 (number of star teachers = ½ total number of teachers) for both designs

1. What Are the Study’s Main Outcomes and How Should They Be Measured?

We recommend that the demonstration use school records data on student achievement and student background, administer a school screening survey, and administer a teacher questionnaire in the fall of the program year, with at least one year of followup. Additional data collection activities may be added at the suggestion of IES or experts consulted for refinement of the study design.

IES has stated that the main question of interest is the impact on student achievement. To this end, it is important to collect student test score data that can be linked to teachers participating in the study. To reduce costs and minimize burden on participating schools and districts, we recommend the use of routinely administered district assessments rather than new researcher-administered tests. The study can exploit existing school records data if student scores can be linked to teachers, linked by student from one year to the next, and linked by ID code to other student background data such as demographic characteristics, special education status, attendance, and mobility. We assume that such rich administrative data will be available; the data were a requirement for inclusion of the districts in the study (thus permitting the demonstration to estimate value-added indicators for screening star teachers).
IES has also expressed an interest in learning about the impacts of the star teacher program on recruitment and retention. Recruitment outcomes can be measured by identifying vacancies at some point before the start of the school year and then examining the qualifications of the staff who fill those vacancies. Qualifications can include teachers’ certification status, preparation route, examination scores (such as the Praxis II series subject matter examinations that many states require), coursework in subject area, subject-certification, years of experience, or attainment of advanced degrees. Simple measures can be constructed, such as the rate of out-of-field teaching, by applying the criteria used by the state for in-field teaching, for example, to define highly qualified teachers (HQTs).

Retention outcomes are straightforward to measure. Investigators merely have to track down study participants one year or more after the beginning of the study and verify their employment status. The evaluation can focus on measuring retention in the school, in the district, or in the profession during the first school year, after one year, during the second year, after the second year, and/or beyond. Administrative records can also be used to track teacher progress if they are available in sufficient detail for research and if follow-up teacher surveys are deemed too costly. Follow-up surveys require a field data collection effort, but they allow the researcher to probe in more detail about the nature of the teachers’ current status, the dates of any transitions, and other factors related to career mobility.

2. What Grades and Subjects Should the Study Include?

We recommend focusing on middle school mathematics (e.g. grades 7 and 8) if a between-school design is adopted, and elementary reading and mathematics (grades 4 through 6) if a within-school design is adopted.

Students in younger grades should only be included if there is a pretest available (e.g. grade 2 as a pretest for grade 3), allowing the researcher to construct gain scores or value added scores.

Data availability and study design play a major role in the grades and subjects that can be included in the study. Ideally, the demonstration would focus on the teaching assignments for which there is the greatest demand, such as secondary mathematics, science, or special education. In practice, the choice of grade level depends on the unit of random assignment. If students are randomly assigned to teachers within schools (Design A), studying middle or high school grades would be impractical, because the structure of classes hinders a comparison of teachers within a grade; at the secondary level, classrooms within a grade are typically sorted according to ability level—rarely do two teachers in a given grade who teach the same subject also teach students of the same ability level. Even when two teachers in a given grade do teach the same subject and ability level, it is likely that scheduling conflicts will impede random assignment. On the other hand, it is possible with the between-school design (Design B) to select any teaching assignment in the school, as long as it is possible to identify many schools with a vacancy in the same grade and subject.

The other constraint on grades and subjects that can be included in the study is the availability of standardized tests. Routinely tested grades and subjects include mathematics and reading in grades 3 through 8. We note that middle school mathematics is the only
routinely tested grade and subject that represents a critical shortage area in many teacher labor markets, and therefore is the most desirable position to target if a between-school study design is deemed feasible.

3. What Other Data Would Be Useful?

In addition to outcome data, background data on the students, teachers, and schools would strengthen the study, as would information on process and program implementation. Some background data can be compiled from administrative records, but new data collection designed specifically for the study could most effectively and accurately measure the phenomena of interest.

**Student data.** School records usually contain data on students’ race/ethnicity, free lunch program participation status, disability status (Individualized Education Plan), and, possibly, enrollment and attendance. Assuming that the data can also be linked to earlier test scores, these data are probably sufficient for the analysis required in the study.

**Teacher data.** The evaluation would most likely require a new survey to provide data on teachers. The proposed study would need information on teachers’ experience in the teaching profession, years of teaching in the study schools, teacher preparation and general education background, certification and degrees, and demographic characteristics. The administrative records that districts maintain on teachers are usually inconsistent, incomplete, outdated, or inaccessible. Our experience with many other large-scale education studies suggests that teacher variables are most easily measured with a teacher survey, such as a self-administered written questionnaire with telephone followup.

**School data.** School data could be compiled from both administrative and survey data. The Common Core of Data (CCD), produced by the National Center for Education Statistics (NCES), is a rich source of information on schools and districts throughout the country, but the information is typically released with a two- to three-year lag. A promising strategy might be to use the CCD to identify promising districts but work with districts to construct a school screening tool—for example, a telephone interview protocol—that would ask the office staff in each school about key features such as the following: staff turnover (possibly verified by obtaining faculty rosters), classroom organization (such as ability grouping or subject compartmentalization), classroom aide assignments, mentoring and professional development resources, and experience and other qualifications of the school’s principal and assistant or vice principal(s).

**Process data.** Process variables would be especially useful for gauging the success of the star teacher recruitment and documenting the process for future replication (if warranted). For example, researchers should keep track of the number of teachers initially targeted, the number of targeted teachers who applied to the program, the number of applicants screened (including any numeric scores or rankings if scores/rankings were to be used in evaluating and selecting star teachers), the number of participants selected and the number successfully matched to schools, the number of matched participants who completed the program, and the number who returned for a second year. Little is currently
known about the rates at which teachers might drop out of the recruitment pipeline and the points along the way at which dropout would be greatest. To the extent possible, the evaluation should document the characteristics of teachers who progress to each stage and the barriers to reaching each stage.

In addition, the evaluator should attempt to document the ways in which teaching slots from the sending schools were filled when the star teachers vacated them. This would provide some data on the possible adverse effect of the star teacher program on sending schools, which may be a concern to districts thinking about implementing the program in the future.


