An Impact Evaluation of the Teacher Incentive Fund (TIF)

Part B

July 13, 2011

Contract Number:
ED-04-CO-0112 (0012)

Mathematica Reference Number:
06715-400

Submitted to:
Institute of Education Sciences
IES/NCEE
U.S. Department of Education
555 New Jersey Avenue, NW
Washington, DC 20208
Project Officer: Elizabeth Warner

Submitted by:
Mathematica Policy Research
600 Maryland Avenue, SW
Suite 550
Washington, DC 20024-2512
Telephone: (202) 484-9220
Facsimile: (202) 863-1763
Project Director: Jill Constantine

MATHEMATICA
Policy Research, Inc.

CONTENTS

TABLES

SUPPORTING STATEMENT FOR PAPERWORK REDUCTION ACT SUBMISSION

This package requests clearance from the Office of Management and Budget (OMB) for data collection activities to support a rigorous evaluation of the Teacher Incentive Fund (TIF). This evaluation will include TIF grantees who were awarded funds from the American Recovery and Reinvestment Act (ARRA) of 2009 and the U.S. Department of Education's (ED) fiscal year (FY) 2010 appropriation. The Institute of Education Sciences (IES), within ED, has contracted with Mathematica Policy Research and its partners Chesapeake Research Associates and faculty and staff at the Peabody College of Education at Vanderbilt University to conduct the evaluation.

The main objective of the evaluation is to estimate the impact of differentiated performance-based incentive pay (DPBIP)[1] on student achievement and the mobility and retention of teachers and principals. The evaluation design is an experiment in which researchers will randomly assign schools within a district to either a treatment or control group. The treatment schools will implement educator DPBIP as part of a performance-based compensation system (PBCS). Control schools will implement the same non-differentiated components of the PBCS program and a one percent across-the-board bonus, but will not implement any type of DPBIP for the duration of the TIF grant. We will compare student achievement and other outcomes between the treatment and control schools to estimate the impact of DPBIP compared to the one percent bonus.

The Notice of Final Priorities (NFP) for the TIF grants, published in the *Federal Register* on May 21, 2010, announced two competitions for grants to be awarded in 2010—the TIF main competition and the TIF evaluation competition; applicants applied to one or the other competition. Successful applicants for the evaluation competition received an "evaluation grant" that includes an additional financial award to fund TIF program activities, including some activities that are not eligible for funding under the main competition.[2] Grantees awarded an evaluation grant had to demonstrate their ability and willingness to meet the grant requirements, which included the main competition requirements plus additional ones specific to the evaluation. In particular, evaluation grantees agreed to cooperate with data collection activities required for the national evaluation, identified the schools that will participate in the national evaluation, and agreed to allow those schools to be randomly assigned to either the treatment or control group. Both main and evaluation grants are for five years.

This is the second submission of a two-stage clearance request for the evaluation. The first package (approved October 18, 2010, under OMB Control Number 1850-4285) requested clearance to ensure that grantees' program designs and implementation are consistent with the requirements for a rigorous evaluation of the TIF, and if necessary, recruit grantees for the evaluation. This second package requests clearance to collect data that will support the full-scale study.

[1] For this document, DPBIP refers to the differentiated incentive pay portion of a grantee's PBCS. DPBIP programs provide bonuses for highly effective teachers and principals, where effectiveness is based on student achievement growth, observations, and any other criteria included in the district's PBCS.

[2] Evaluation grantees will receive $250,000 per pair of schools identified for the national evaluation, up to $2.5 million for 20 schools.

We believe it is important to note that our eventual data collection plans will differ in two ways from those for a study of TIF grantees being conducted by the Policy and Program Studies Services (PPSS) in the Office of Planning, Evaluation and Policy Development at ED. First, the two data collection efforts target different respondents. The PPSS study includes grantees from the FY2007 awards while participants in the current study received their grants in FY2010, and the two studies target *different* schools and/or educators. Second, the focus and design of each study is different. The PPSS evaluation is an implementation study. This evaluation uses a rigorous experimental design in which schools are randomly assigned to either a control or a treatment group to estimate the impact of DPBIP on student achievement and educator mobility and recruitment.

Part B: Collection of Information Employing Statistical Methods

1. Respondent Universe and Sampling Methods

This study sample will rely on a convenience sample that consists of grantees that were awarded an evaluation grant and include the schools these grantees identified for the evaluation and the educators working at these schools. The study will not statistically sample 2010 TIF grantees. The TIF grant competition provided applicants with an option to apply for an evaluation or a main grant. In addition to meeting other requirements, the evaluation grantees must be willing to allow at least eight high-need schools with tested grades within a district to be randomly assigned to either a treatment or control group. ED awarded 11 TIF evaluation grants covering 15 districts[3] and over 250 schools. In order to obtain estimates of the effect of the DPBIP program at the desired level of precision, we need to include approximately 250 schools in the evaluation (see Table 1 below).

The proposed data collection will include the following components:

- **District survey.** We will survey all 2010 TIF districts (from main and evaluation grantees) to determine specific features of the incentive program, to understand approaches districts used to obtain buy-in and compromises they had to make, explore districts' experiences, and compare characteristics of main and evaluation districts. We will also use the data on evaluation district programs to examine the association between impacts and key program features.
- **District interview.** The questions in the district interviews will allow us to collect more in-depth information on evaluation district programs than that collected from the survey, and to probe for clarification if necessary. We will use this detailed information to more thoroughly understand each program's context, implementation strategy, and challenges. All evaluation districts will be interviewed.
- **Principal and teacher administrative data.** These data will be used to estimate the impacts of DPBIP on educator mobility and recruitment. The data will also allow us to examine the association between educator characteristics and student and educator outcomes, and to describe the educator sample. These data will be collected for all principals and teachers in study schools.

[3] New York City is a grantee and is also included in the New York State grant.

- **Student records data.** We will use existing state or district test score data to estimate the impact of DPBIP on student achievement, the key outcome of interest. Information on students' demographic and socioeconomic characteristics and their achievement test scores prior to the study school year will be used to describe the students in the study and to develop more precise impact estimates. To the extent possible, we will use student-teacher linked data to estimate teachers' value-added score to better understand mobility of high- and low-performing educators. We will collect math and reading scores for all third through eighth grade students in the study schools. We will also collect information from participating districts on the teacher performance and payouts under the TIF program.
- **Principal survey.** The principal survey will be used to assess hiring practices, classroom assignments, knowledge and perceptions of the TIF program in the study schools, how this may change over time, and to supplement administrative data to be obtained from district records. The principal survey can also provide important insight on their motivation for remaining, leaving or entering a study school. We will survey all principals of study schools.
- **Teacher survey.** The teacher survey will be used to assess knowledge and perceptions of the TIF program in the study schools, how this may change over time, and to supplement administrative data to be obtained from district records. The teacher survey can also provide important insight on teachers' motivation for remaining, leaving, or entering a study school. The survey will be administered to a sample of approximately 2,000 teachers from the anticipated sample of 250 schools in the evaluation.

The evaluation does not aim to make statements that generalize beyond the districts and schools under study. Although we will not be able to generalize findings to all 2010 TIF grantees, we will obtain valid estimates of the impacts for the set of districts in our study. We will also be able to describe the characteristics and implemented policy of the evaluation and non-evaluation districts from the FY10 competition.

2. Procedures for the Collection of Information

a. Statistical Methods for Sample Selection

As described above, the study will use a convenience sample of TIF evaluation grantees, and will comprise as many schools as the evaluation districts are willing and able to include, up to the maximum of 20 allowed. Moreover, the district survey will be administered to all TIF districts covered by either main or evaluation TIF grants. Thus, the study will not statistically sample grantees, districts, principals, schools, or students. We will sample teachers for the teacher survey as explained below.

The teacher survey will be administered to a representative sample of teachers in the study schools. We plan to survey two types of teachers—those in tested grade/subject combinations and teachers in nontested grade/subject combinations. This will be done in two grade spans—elementary and middle school grades. We need adequate representation of both tested and nontested grade-subjects because each group of teachers faces different incentives and we need to measure the impacts that performance-based pay has on each group separately. Those in nontested grade/subjects may be subject to bonuses based on performance measures that they only indirectly affect, whereas teachers in tested grade/subjects have a more direct effect on performance measures. We will focus separately on elementary and middle schools. At the elementary level, the teachers are typically in nondepartmentalized settings, meaning that the classroom teacher is responsible for all subjects (including both math and reading), whereas in middle school the teacher is typically responsible for one subject (such as math *or* reading).

We plan to draw a census of all teachers who are responsible for math or reading instruction in all study schools in grades four and seven. By sampling teachers in this manner, we can limit the amount of heterogeneity of teachers that could make it difficult to interpret the treatment effect. For example, in several analyses, teacher survey data will be linked with test score data to examine DPBIP impacts on differential mobility of high- and low-performing teachers. If all of our elementary teachers come from a particular grade (and likewise for middle school teachers), then the test scores used in this analysis will already be comparable between the treatment and control groups within each district. On the other hand, if we sample teachers from several grades, sample sizes of teachers per grade will be somewhat small, and there may be the possibility of grade imbalance (among either teachers or students) between the treatment and control groups. We will then have to rely on inclusion of grade dummies and other modeling approaches to take care of this potential grade imbalance. Overall, we believe it is preferable to limit the generalizability of our findings to avoid having to adjust for possible grade imbalances.

We anticipate an average of four teachers per elementary school, six per middle school (three each in math and reading), and eight per K-8 school (or school with both elementary and middle school grades). This is a total of 1,300 teachers (4 x 150 elementary + 6 x 50 middle school + 8 x 50 K-8 schools).

For nontested grade/subjects, we will focus on first grade teachers in elementary schools and grade seven science teachers in middle schools. We selected the first grade because we expect that every elementary school will have a full-day class and is less likely to have standardized testing than grades two and three. We selected science because it is a well-defined subject that is not routinely tested, but for which retaining certified teachers is an important policy goal.

We will randomly select two first grade teachers from every school with elementary grades, two grade seven science teachers from every school with middle school grades, and two of each teacher from schools with both elementary and middle school grades. We anticipate this would result in a sample of 300 teachers from elementary schools, 100 from middle schools, and 200 from K-8 schools, for a total of 600 teachers. Therefore, the total number of teachers we plan to survey is approximately 2,000.

The teacher survey data is more useful if we are able to follow the same teachers over time as well as refresh the sample with teachers who are new to the TIF schools, to learn about the types of teachers each school is recruiting. Therefore, our sampling plan for the follow-up surveys is to survey 100 percent of leavers, 100 percent of replacement teachers, and approximately 75 to 82 percent of stayers. If we assume that 15 to 20 percent will leave the sample, and an equal number replace them, then we must reduce the sample of stayers by 18 to 25 percent, or in other words, survey 75 to 82 percent of the stayers. This is equivalent to following the 285 to 380 expected leavers, adding a roughly equal number of replacement teachers, and following 1,140 to 1,330 randomly chosen stayers.

In subsequent years we will continue to survey all leavers and their replacements, since they are a smaller group and are of great policy interest, and follow a shrinking percentage of the stayers. If the exit rates after each year are higher than we have anticipated, we will consider altering the sampling rates by mobility status in order to achieve adequate sample sizes in each group.

b. Estimation Procedures

Random assignment of schools within a district to a treatment group that will implement DPBIP or to a control group not allowed to do so for the duration of the TIF grant is an ideal design for assessing overall effectiveness. Our primary impact analysis will exploit this experimental design to provide rigorous estimates of the impact of DPBIP on student achievement and teacher/principal mobility and recruitment. Additional nonexperimental analyses are designed to estimate the relative effectiveness of individual-based versus group-based or mixed incentive programs, explore the association of other key program features with student achievement and teacher/principal outcomes, and to learn about districts' implementation experiences and challenges.

Estimating the overall impact of DPBIP. With this experimental design, the simple differences between mean outcomes in the treatment and control schools should yield unbiased estimates of the impacts of DPBIP. However, the precision of the estimates can be improved by using regression procedures to control for student, teacher, or school baseline characteristics that may explain some of the variation in outcomes not related to the treatment itself. These characteristics may include student controls, such as test scores from the year before TIF implementation; gender, race/ethnicity, free- or reduced-price lunch eligibility, special education status, and English learner status; teacher controls, such as demographic characteristics, age, experience, and educational background; and school-level averages of the student or teacher characteristics. Regression procedures also enable us to adjust for any differences between treatment and control groups in these baseline characteristics that happen to arise due to chance or sample attrition. The regression model must be flexible enough to include the full range of programs and generate estimates of district-specific impacts, which can then be aggregated to produce an overall estimate. We will therefore estimate variations of the following model for the outcome y_{ijk} of individual (student or teacher) *i* in school *j* within district *k*:

$$(1) \quad y_{ijk} = \mathbf{R_{jk}}\boldsymbol{\alpha} + \sum_{k=1}^{K} \beta_k (T_{jk} \times G_k) + \mathbf{X_{ijk}}\boldsymbol{\delta} + \mathbf{Z_{jk}}\boldsymbol{\gamma} + u_{jk} + \varepsilon_{ijk}$$

where $\mathbf{R_{jk}}$ is a vector of indicators for combinations of grade levels and randomization strata; $\boldsymbol{\alpha}$ is a vector of grade-by-strata fixed effects; T_{jk} is a treatment indicator; G_k is a dummy variable for district *k*; β_k is the impact of DPBIP in district *k*; $\mathbf{X_{ijk}}$ is a vector of baseline individual characteristics with coefficient vector $\boldsymbol{\delta}$; $\mathbf{Z_{jk}}$ is a vector of baseline school-level characteristics with coefficient vector $\boldsymbol{\gamma}$; u_{jk} is a random school effect; and ε_{ijk} is a random individual error term. The district-specific impacts of performance pay, β_k, are the key coefficients of interest in equation (1). We will estimate equation (1) with ordinary least squares (OLS) using Huber-White ("sandwich") standard errors that account for school-level clustering.

Our primary interest is in the overall, average impact of DPBIP in the full study sample. To estimate the average impact of DPBIP on schools in the study, we will take a weighted average of the estimated district-specific effects, $\hat{\beta}_k$, with weights equal to the number of treatment and control schools within each district. The standard error of the average impact estimate can be calculated from the estimated variances and covariances among the district-specific impacts from equation (1).

The evaluation includes four years of analyses. Impacts in the second and subsequent years of the implementation of the DPBIP may be larger than those in the first year for several reasons. First, changes in educator effectiveness and the composition of the teaching staff at treatment schools may be more pronounced after educators observe the payments from earlier years. Also, if educators improve their performance over time, in years 2 through 5 of the grant, some students will have had multiple years of exposure to the treatment. For these reasons, equation (1) will be estimated separately for assessing impacts for each year of implementation, as well as cumulative impacts.

The impact of DPBIP on the outcomes of interest—student achievement and educator mobility and recruitment—will be estimated with a variant of equation (1). Student achievement outcomes are math and reading scores from spring 2012, 2013, 2014, and 2015 state or district assessments. Because tests will differ across states, grade levels, and subjects, we will convert raw scale scores to z-scores (raw scores minus the mean score divided by the standard deviation of scores on that test among students in that grade and state) in order to scale the outcome variable comparably across all students in the sample. Using district records, we will measure teacher and principal retention as a dichotomous outcome for whether or not the teacher returns to work in the grantee site and/or in his or her initial school in fall of 2011 and continue to do so annually through 2015. Because the retention outcome is dichotomous, we will estimate the probit model analog of equation (1). Annual school-level teacher data from study schools in fall, 2011 through fall, 2015 (from district records) and spring 2012, 2013, 2014, and 2015 (from the principal and teacher surveys) will be analyzed as outcomes to examine impacts on the composition of the teaching staff. If available from administrative records, the quality of applicants who apply to teach in study schools for school years 2012–2013, 2013–2014, 2014–2015, and 2015–2016 will also be analyzed, including the total number of applicants, average experience level, percentage of applicants who have teaching experience, and the selectivity of the college from which they graduated. Equation (1) can be aggregated to the school level for the analysis of composition outcomes.

To better understand mobility of high- and low-performing principals and teachers, for districts where we can obtain or calculate a measure of staff effectiveness, we will also estimate a model of transitions that includes a teacher or school measure of effectiveness, and interactions of this measure with treatment indicators in the set of independent variables. The coefficients on the effectiveness measure by treatment interactions provide an estimate of whether differences in retention between highly effective and less effective principals or teachers are more or less pronounced in treatment versus control schools. Since high- and low-performing teachers are not being randomly assigned to treatment and control schools, and estimates of their effectiveness may be endogenous if DPBIP induces greater teacher effectiveness, these estimates are nonexperimental and will need to be interpreted with caution. Wherever possible, we will obtain or calculate value-added estimates based on student achievement to measure teacher effectiveness. In addition, if possible, we will also use districts' measures of effectiveness.

Estimating the effectiveness of key program features. We will conduct exploratory analyses to assess whether particular features of DPBIP are associated with impacts on student achievement. These analyses will, in particular, examine the relative effectiveness of DPBIP models that place different weights on individual versus group performance in the determination of incentive payouts. Other programmatic features of interest include the average and maximum size of the incentive payouts and the degree to which the payouts vary across educators.

Since we do not expect that districts will randomly assign specific components of their DPBIP to schools, we will not be able to experimentally assess the relative effectiveness of different DPBIP program features. Instead, we will examine the association between impacts and key program features in a regression framework. We will be careful to note that an observed association between impacts and programmatic features may not necessarily have a causal interpretation.

For these analyses, we will rely on findings from the implementation analysis to examine how the variation in programmatic features is related to the impact. Our basic approach is to regress the estimated district-specific impacts from equation (1) on a measure of a specific programmatic feature. For the estimated impact $\hat{\beta}_k$ from district *k*, we estimate:

$$(2) \quad \hat{\beta}_k = \pi_0 + \lambda W_k + \omega_k$$

where π_0 is an intercept, W_k is a measure of a specific programmatic feature with associated coefficient λ, and ω_k is an error term that includes random error in estimating the true impact β_k. Because impacts might be more precisely estimated in some districts than in others, we will weight districts by the precision of the estimated impacts when estimating equation (2) to account for this source of heteroskedasticity in the error term. For each of the programmatic features described earlier, we will estimate equation (2) with the specified program feature as the only covariate, given the limited number of grantees in the sample.

Understanding the implementation experiences of TIF districts. Understanding the implementation experiences and challenges of TIF districts will provide essential information for documenting their incentive policies and experiences and for improving the implementation of future incentive programs. The descriptive analyses will describe the average characteristics of three groups of TIF districts: (1) evaluation districts; (2) main districts; and (3) the combined population of all TIF districts. Since the evaluation grantees were purposively selected, and the impact estimates cannot necessarily be generalized beyond this sample, we will use the district surveys to construct tables on their incentive policies, comparing the evaluation districts to all recent district awardees. We also will use the district surveys and information from telephone interviews (for the evaluation districts) to document and analyze implementation challenges.

We will also analyze the implementation data collected from evaluation grantee, district, and school documents; district, principal, and teacher surveys; and telephone interviews to provide crucial context for the interpretation of the impact findings. The principal and teacher surveys will provide context to determine if they understood the incentive compensation policy and program in their district and school and adjusted their behavior accordingly. After the initial survey, for each subsequent wave of the principal and teacher surveys, we will construct tables to assess any changes in educators' understanding and behavior.

Comparing the outcomes for TIF districts to non-TIF districts. In addition to estimating the impact of the DPBIP, we will plan to tabulate outcomes for a group of TIF schools that includes both treatment and control group members, and a reference group of non-TIF schools that are not implementing any kind of PBCS. The goal of this analysis is to provide information on the broader set of TIF-funded reforms beyond performance pay. Outcome data for non-TIF schools, such as average test scores, and PBCS implementation status will be obtained from publicly available data sources.

Degree of accuracy needed. The study must ensure adequate statistical power for detecting policy-relevant impacts on key outcomes. The study is powered to detect a minimum detectable effect (MDE) of 0.09 of a standard deviation on student achievement (see Table 1).

Table 1. Minimum Detectable Effects on Student Test Scores

	Number of Schools	Number of Students	Minimum Detectable Effect on Student Test Scores
Proposed sample (MDE = 24 percent annual gain)	250	89,000	0.09
Sample to detect MDE = 22 percent annual gain	310	110,360	0.08

The calculations in Table 1 assume the following: (1) 80 percent power and a 5 percent significance level for a two-tailed test; (2) 60 percent of schools in the sample will be elementary schools, 20 percent middle schools, and 20 percent K-8; (3) each elementary school will contain 240 students in tested grades, each middle school will contain 740 students in tested grades, and each K-8 school will contain 320 students in tested grades; (4) test scores will be missing for 15 percent of students in tested grades and 13 percent of the total variance of student test scores will be between schools; and (5) covariates can explain 65 percent of the between-school variance and 40 percent of the within-school variance of student test scores in middle schools; covariates explain only 50 percent of the variance between elementary schools and 33 percent within elementary schools, and explain 55 percent of the variance between K-8 schools and 35 percent within K-8 schools.

There is no universally agreed upon MDE that education interventions should be powered to detect. One strategy is to put the MDE in context by expressing it in terms of expected annual learning growth of students. Recent work by Bloom et al. (2008) indicates that the expected annual growth in student test scores is different for reading and math, and both decline as student grade levels increase. However, the mean annual growth for reading and math for students in third through eighth grades is approximately .37 standard deviations. Therefore, the MDE for this study, with our proposed sample size, is powered to detect a difference of approximately 24 percent of a year, roughly 2.9 months, of learning. To detect an MDE of .08, or roughly 22 percent of a year of learning, the evaluation would require 310 schools, shown in the bottom row of Table 1.

DPBIP programs are theorized to improve student achievement through educator mobility and refining teaching practices. Since both of these mechanisms will take time to be realized, effects in the first year or two may not be detectable. In addition, students may realize the effect each year they experience a high performing teacher, and small yearly effects may accumulate over a number of years. This study is powered to be able to detect impacts that might occur in a single year, but is more likely to detect impacts that accumulate over multiple years.

The study has also been powered to detect teacher response to DPBIP. One key outcome is the percentage of teachers who are retained in the school each year, which is obtained from the teacher survey.[4] Based on national statistics (Ingersoll 2003), we expect the annual teacher retention rate in the control schools to be between 80 and 90 percent. A survey of the full sample of teachers would be unnecessarily burdensome without providing substantial gains in statistical power. Therefore, we will select a sample of 1,900 teachers, as discussed above. This will enable us to detect approximately

[4] If possible, we will also use district records to estimate teacher retention; we expect that the quality of the district records will vary substantially across districts.

a six percentage point impact, as illustrated in Table 2, on the basis of the same assumptions as those from the student-level MDE calculations. This sample size also allows sufficient power to detect meaningful impacts in a 50 percent subgroup of teachers, in order to examine, for example, effects for novice and more experienced teachers.

Table 2. Minimum Detectable Impacts on Teacher Retention

	Number of Schools	Number of Teachers	Minimum Detectable Percentage Point Change if Control Group Retention Rate is:	
			90%	80%
All teachers in sampling frame	250	3,500	4.0	5.3
Proposed sample	250	2,000	4.9	6.6
50 percent subgroup of proposed sample	250	950	6.5	8.7

3. Methods to Maximize Response Rates and Deal with Nonresponse

There are multiple strategies to maximize response while minimizing burden on respondents, and the following techniques are major contributions to a high completion rate: establishing positive relationships with respondents and school and district staff; sending letters prior to the surveys; and establishing efficient and flexible scheduling. We will include a statement on confidentiality and data collection requirements (Education Sciences Reform Act of 2002, Title I, Part E, Section 183) in all letters, data collection instruments, and study brochure (described in detail in Part A and appendices). In cases when the data collection activity is voluntary (e.g. the teacher survey), we will include a statement indicating that participation is voluntary, yet emphasize the importance of their response for the study findings.

We anticipate a district response rate of at least 80 percent overall (from both main and evaluation grantees). We expect complete cooperation from evaluation districts because evaluation grantees committed to cooperating with data collection activities for the national evaluation, which is a condition of the grant. Furthermore, during the implementation year, the research team will be working with evaluation grantees and will establish a rapport with them. To further solidify administrators' cooperation, we will adhere to additional data collection requirements that districts may have such as preparing research applications and seeking institutional review board (IRB) approvals. In our correspondence, we will also send notification letters on ED letterhead to districts to capture their attention and to help increase the response rate.

Based on Mathematica's experience in conducting surveys with teachers and principals, we expect at least an 85 percent response rate for the principal and teacher surveys. Because principals signed letters of commitment to the evaluation, we anticipate this will help enhance cooperation from both teachers and principals. To ensure response, first a follow-up will be initiated through email and telephone calls to educators who do not respond within two to three weeks of the initial contact. Second, nonrespondents will be given the option of providing data during the telephone follow-up. Data collectors can read the questions aloud and enter responses on the hard copy instrument or directly into the web-based survey. Third, experienced interviewers will be recruited and extensively trained on data collection procedures, including methods for promoting cooperation among school staff. Interviewers especially skilled at encouraging cooperation will be available to persuade reluctant educators to participate.

We thoroughly pretested the instruments for such features as clarity, accuracy, length, flow, and wording. Trained quality control staff checked for completeness and reasonableness as soon as a hard copy questionnaire was received and followed up with respondents if problems are identified. The web-based survey will not allow respondents to enter out-of-range or inconsistent responses, and data entry programs will also check for these.

Reducing districts' burden in the submission of study data will facilitate attaining a response rate of at least 85 percent on student records and educator administrative data. Federal rules permit ED and its designated agents to collect student demographic and existing achievement data from schools and districts without prior parental or student consent (Family Educational and Rights and Privacy Act (FERPA) (20 U.S.C. 1232g; 34 CFR Part 99)). To maximize the response rate and minimize burden on schools and parents, we will follow these federal rules.

Finally, we will be courteous but persistent in our follow-up with participants who do not respond quickly to our attempts to reach them.

4. Tests of Procedures or Methods to be Undertaken

As much as possible, data collection instruments for the study drew upon surveys and protocols that have been used successfully in previous studies. The pretests assessed the content and wording of individual questions, organization and format of the questionnaire, respondent burden time, and potential sources of response error. We piloted instruments that are new, that are adaptations and extensions of existing ones, that have limited information on reliability and validity for the population in this study, and for which we wish to examine how measures perform when combined with others. .

Each of the educator surveys was pretested with no more than nine respondents, to identify problems future respondents might have in providing the requested information. Responses on the survey were collected from educators by email, regular mail or fax, and the study team debriefed with respondents by telephone or email. The district survey and interview protocol were pretested with representatives from districts that received TIF grants from an earlier round. The results of the pretests were used to revise and improve the instruments.

5. Individuals Consulted on Statistical Aspects of the Design, Data Collection, and Data Analysis

The following individuals were consulted on the statistical aspects of the study:

Name	Title	Telephone Number
Jill Constantine	Associate Director of Research and Education Area Leader, Mathematica	609-716-4391
Steven Glazerman	Senior Fellow, Mathematica	202-484-4834
Matthew Springer	Assistant Professor of Public Policy and Education; Director, National Center on Performance Incentives, Vanderbilt University	615-322-5524
John Deke	Senior Researcher, Mathematica	609-275-2230

Alison Wellington	Senior Researcher, Mathematica	202-484-4696
Dan Player	Academic Director, Partners for Leadership in Education, Curry School of Education, University of Virginia	434-243-0904
Hanley Chiang	Researcher, Mathematica	617-674-8374

The following individuals will be responsible for data collection and analysis:

Name	Title	Telephone Number
Jill Constantine	Associate Director of Research and Education Area Leader, Mathematica	609-716-4391
Sheila Heaviside	Associate Director of Survey Research, Mathematica	202-484-3096
Steven Glazerman	Senior Fellow, Mathematica	202-484-4834
Matthew Springer	Assistant Professor of Public Policy and Education; Director, National Center on Performance Incentives, Vanderbilt University	615-322-5524
Annette Luyegu	Survey Researcher, Mathematica	202-264-3463
Alison Wellington	Senior Researcher, Mathematica	202-484-4696
Hanley Chiang	Researcher, Mathematica	617-674-8374

REFERENCES

Bloom, Howard S., Carolyn J. Hill, Alison Rebeck Black, and Mark W. Lipsey. "Performance Trajectories and Performance Gaps as Achievement Effect Size Benchmarks for Educational Interventions." MDRC Working Papers on Research Methodology. New York, NY: MDRC, October 2008.

Ingersoll, Richard M. "Is There Really a Teacher Shortage? A Research Report." Center for the Study of Teaching and Policy, University of Washington, 2003.

MATHEMATICA
Policy Research, Inc.

www.mathematica-mpr.com

Improving public well-being by conducting high-quality, objective research and surveys

Princeton, NJ ■ Ann Arbor, MI ■ Cambridge, MA ■ Chicago, IL ■ Oakland, CA ■ Washington, DC