An Impact Evaluation of the Teacher Incentive Fund (TIF)

Section B

July 21, 2010

Revised October 12, 2010

MATHEMATICA
Policy Research, Inc.

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

An Impact Evaluation of the Teacher Incentive Fund (TIF)

Section B

July 21, 2010

Revised October 12, 2010

CONTENTS

TABLES

PART B: SUPPORTING STATEMENT FOR PAPERWORK REDUCTION ACT SUBMISSION

This OMB package requests clearance to ensure that grantees' program design and program implementation are consistent with the requirements for a rigorous evaluation of the Teacher Incentive Fund (TIF), and if necessary, recruit grantees for the evaluation. This evaluation will include TIF grantees who are awarded funds from the American Recovery and Reinvestment Act (ARRA) of 2009 and the U.S. Department of Education's (ED) fiscal year 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 teacher and principal (hereafter, educators) mobility and retention. 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 1% across-the-board bonus but will not implement any type of DPBIP throughout the duration of the TIF grant. We will compare student

[1] For this document, DPBIP refers to the differentiated incentive pay portion of a grantee's performance-based compensation system (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.

achievement and other outcomes between the treatment and control schools to estimate the impact of DPBIP compared to the 1% bonus.

The Notice of Final Priorities (NFP) for the TIF grants, published in the *Federal Register* on May 21, 2010, proposed two competitions for grants that will be awarded in 2010—the Main TIF competition and the TIF Evaluation competition; applicants apply to one or the other competition. Unsuccessful applicants for the evaluation grant will automatically be considered for the Main TIF competition. Successful applicants for the Evaluation competition will receive an "evaluation grant" that includes an additional financial award to fund TIF program activities, including for some uses that are not eligible for funding under the Main competition.[2] Grantees awarded an evaluation grant must demonstrate their ability and agreement to meet the grant requirements, which includes the Main competition requirements plus additional ones specific to the evaluation. Even so, we anticipate that we will need to work with grantees to confirm the requirements of the evaluation and to ensure their successful participation.

This is the first of two requests for the evaluation. A future request will seek clearance to collect educator and student records from districts, administer grantee and educator surveys, and conduct grantee interviews. We are submitting the package in two stages because ensuring that grantees' program design and program implementation are consistent with the requirements of the evaluation must begin before all the data collection instruments are developed and pretested. Also included in this first request

[2] The NFP states an evaluation grantee will receive, minimally, an extra $1 million, and can receive as much as $2 million.

is the draft letter to participating districts and principals (Appendix A), an information sheet that will be included with the district/school letter (Appendix B), Mathematica's internal confidentiality pledge (Appendix C), and topics to be discussed and goals of the initial and follow-up phone conversations and site visits that will occur shortly after grants are awarded (Appendix D).

We provide an overview of the study's eventual data collection plans in order to provide context, but they are not the focus of this request. We believe it is also important to note that our eventual data collection plans will differ from those for a study on TIF grantees being conducted by 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 will receive their grants in FY2010. Second, the focus and design of each study is different. The PPSS evaluation is an implementation and feasibility study. Its aim is to describe grantees' program features and implementation experiences, as well as examine the feasibility of using extant data to examine the association between TIF participation and student achievement and educator outcomes. This evaluation uses a rigorous experimental design in which schools are randomly assigned to either a control or treatment group to estimate the impact of DPBIP on student achievement and educator mobility and recruitment. For these reasons, the data collection requirements for this evaluation differ from the current PPSS study.

COLLECTION OF INFORMATION EMPLOYING STATISTICAL METHODS

1. Respondent Universe and Sampling Methods

This study will not statistically sample districts and schools. Instead, it will rely only on districts or other entities that competed for and were awarded a TIF evaluation grant requiring participation in a national evaluation. The evaluation does not aim to make statements that generalize beyond the districts and schools under study. Once evaluation grantees have been notified of their award, we will solidify their participation as described below. We include examples of notification materials and a study description in Appendices A and B.

Selection of Grantees. The TIF grant competition will provide applicants with an option to apply as an evaluation grantee. In addition to meeting other requirements, the evaluation grantees must be willing to allow at least eight schools within a district to be randomly assigned to either a treatment or control group.

In order to obtain estimates of the effect of the DPBIP program at the desired level, we need to include at least 200 schools in the evaluation (see Table 1 below). We anticipate that grantees will include an average of 10 schools in the evaluation, thus we anticipate that 20 grantees will take part in the evaluation. After evaluation grants have been awarded, we will contact the list of evaluation grantees to ensure that grantees' program design and program implementation are consistent with the requirements of the evaluation.

~~If there are fewer evaluation grantees than needed to meet the target sample size, we will search the universe of main competition grantees to identify and contact a pool of main TIF grantees about participating in the evaluation. We anticipate contacting two to three times as many main TIF grantees as will ultimately be needed for the evaluation because they are less likely to meet the evaluation criteria (for example, they may be small districts with fewer than eight schools).~~

~~In order to identify and invite the most promising grantees to participate in the evaluation, we will review the main TIF grant applications and prioritize them according to the following criteria:~~

- ~~**District size.** Grantees should be large enough to include at least eight schools with tested grades and subjects in the evaluation.~~
- ~~**Sufficient data capabilities to support evaluation needs.** Ideally, districts in the sample will be able to provide data on educator characteristics, student demographic information, and test scores that can be tied to individual students and linked to teachers.~~

~~If we need to solicit additional grantees for the evaluation, we will work with IES to determine which ones are best suited for the evaluation.~~

Contacting Grantees. Once the evaluation grantees have been identified, we will send information packets to each district that will be included in the evaluation. The packets will include the following documents (see Appendices A and B):

- **Notification letter.** This letter highlights the importance of learning about the effectiveness of incentive programs, reminds the grantee of their participation in the evaluation, and provides an overview of the study design. The letter will also indicate that a member of the study team will telephone a district representative and provide more details, discuss the district's participation in the

study, and arrange for an in-person meeting with district and school officials.

- **Nontechnical brochure.** This two-page document describes the random assignment process in a simple and nonthreatening way, includes a partial list of evaluation requirements, lists the benefits of participation, and presents data collection activities and timeline. It also identifies the organizations comprising the evaluation team and contact information.

Mailings will be sent via FedEx for quick delivery and to better capture the recipients' attention.

Within two days of the mailing's delivery date, an evaluation team member will telephone the grantee to identify the appropriate contact. In subsequent calls, we will briefly describe the study and confirm the district's agreement to participate. We will also arrange in-person visits with the grantee districts to provide an orientation to the evaluation. At any district-level meetings, we will attempt to involve all pertinent decision makers including key staff involved in the TIF grant application process.

2. Statistical Methods for Sample Selection and Degree of Accuracy Needed

Statistical Methods for Sample Selection. This study will not sample grantees, districts, principals, or students. However, we will administer the teacher survey to a representative sample of teachers in the study schools.

As described in section 1 above, the study will use a convenience sample of grantees that received a TIF evaluation grant, and we will include as many schools as the grantees are willing and able to include in the evaluation, up to the maximum of 16 allowed. If the universe of evaluation grantees brings more than the 200 schools, we will discuss with IES whether or not to include

all schools, which would improve the power and lower the minimum detectable effect for the study.

We will collect administrative records for all students, teachers, and principals in the study. Administrative records will provide information on key outcomes—student achievement and educator mobility and recruitment. We will collect administrative records on all students and educators in the sample to maximize the statistical power of the study. In addition, there is no additional burden for a district to provide records on all students and educators, rather than a sample.

We will administer a survey to all principals in the study. Principals will provide crucial data to examine school leaders' understanding, perceptions, and experiences with DPBIP. In addition, in districts where detailed administrative records on recruitment are not available, principals may be an important source of information on recruitment of teachers in study schools.

We will administer the teacher survey to a representative sample of approximately 2,000 teachers. In addition to providing information on teachers' understanding, attitudes, and experiences with DPBIP, this survey may provide additional information on teacher mobility in some districts with limited administrative records. As explained in the next section, a sample of 2,000 teachers will allow us to detect a meaningful impact of DPBIP on teacher mobility, and reduce the overall burden on teachers compared to administering the survey to all teachers in the study classrooms.

To estimate the impact of TIF on each outcome of interest, we will estimate the following model for the outcome y_{ijk} of individual (student or teacher) *i* in school *j* within grantee *k*:

$$= \sum_{=1} (\quad + {}_{1} ({}^{(1)} \times \quad) + {}_{2} ({}^{(2)} \times \quad)) + \quad + \quad + \quad +$$

where G_k is a dummy variable for grantee *k*; α_k is a grantee fixed effect; $^{(1)}$ $^{(1)}$ and $^{(2)}$ $^{(2)}$ are dummy variables for being assigned to a DPBIP program focused on, respectively, group incentives and individual incentives; $_{1}$ $_{1}$ and $_{2}$ $_{2}$ are grantee-specific impacts of, respectively, group incentive and individual incentive programs (or a mixed model if that is more common); $\boldsymbol{X_{ijk}}$ is a vector of individual baseline characteristics (i.e. if individual, i is a student, X_{ijk} is a vector of student characteristics, and if individual i is a teacher, X_{ijk} is a vector of teacher characteristics); $\boldsymbol{Z_{jk}}$ is a vector of baseline school-level characteristics; **δ** and **γ** are coefficient vectors; μ_{jk} is a random school effect; and ε_{ijk} is a random individual error term. The equation above is estimated with ordinary least squares (OLS) using Huber-White ("sandwich") standard errors that account for school-level clustering.

The evaluation will include 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 effort and the composition of the teaching staff at treatment schools should 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, the equation above will be estimated separately for assessing impacts for each year of implementation, as well as cumulative impacts.

Student outcomes of interest are math and reading achievement from the spring 2012, 2013 2014, and 2015 state or district assessments. Since these differ across states, and scores are not always comparable across grades even in the same state, all student test scores will be standardized using the grade-specific state or district means and standard deviations of the tests. Because a gain of a standard deviation in one grade might not be directly comparable to one in another grade, we will investigate the robustness of the impacts by calculating grade-specific impacts separately and comparing them to the overall impact. The educator-level outcome of interest from district records or the educator survey is a dichotomous outcome for whether or not the educator returns to work in the grantee site and/or his or her initial school in the beginning of the 2012, 2013, 2014, and 2015 school years. Because the outcome is dichotomous, we may estimate a probit model in place of the above equation. School-level teacher data from study schools in fall of 2012, 2013, 2014, and 2015 (from district records) and spring 2012, 2013, 2014, and 2015 (from the educator survey) 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. For the analysis of these school-level composition outcomes, the equation above can be aggregated to the school level.

Degree of Accuracy Needed. The study must ensure adequate statistical power for detecting policy-relevant impacts. The study is currently powered to detect a minimum detectable effect (MDE) of .09 standard deviations on student achievement. Based on the assumption that each grantee will be able to provide an average of 10 schools for the evaluation, we anticipate that we will need a sample of 20 districts for the evaluation.

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= 25 percent annual gain)	200	68,000	.09
Sample to detect MDE= 20 percent annual gain	325	110,500	.07

The calculations in Table 1 assume: (1) 80 percent power and a 5 percent significance level for a one-tailed test; (2) 80 percent of schools in the sample will be elementary and 20 percent middle schools; (3) each elementary school will contain 240 students in tested grades and each middle school will contain 740 students; (4) 22 elementary school teachers per school will be in core subjects in tested grades, as will 25 middle school teachers; (5) 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; (6) 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. Assumptions on the clustering of outcomes and the explanatory power of covariates are based on data from six recent random assignment evaluations in K–12 education. Assumptions on school size are based on tabulations from ED's Common Core of Data (CCD) and the School and Staffing Survey (SASS) databases.

The MDE calculations assume a five percent significance level, as is standard in IES evaluations. However, since the most policy-relevant question is whether DPBIP has a positive impact on educator performance and therefore student achievement, we powered the study based on the ability to detect positive impacts by using a one-tailed test.

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–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 25 percent of a year, roughly 2½ months, of learning.

DPBIP programs are designed 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 more likely it will detect impacts that accumulate over multiple years. To detect an MDE of .07, or roughly 20 percent of a year of learning, the evaluation would require 325 schools, shown in the bottom row of Table 1.

The study has also been powered to detect teacher response to DPBIP. One of the key outcomes is the percentage of teachers who are retained in the school each year, obtained from the teacher survey.[3] 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 representative sample of 2,000 teachers, or an average of 10 teachers per school. This will enable us to detect approximately a 5 percentage-point impact, as illustrated in Table 2. 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.

[3] 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.

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	200	4,478	3.6	4.8
Proposed sample (=10 teachers per school)	200	2,000	4.5	6.0
50 percent subgroup of proposed sample	200	1,000	5.8	7.8

The teacher-level power calculations use the same assumptions as the student level calculations except that the teacher calculations use a two-tailed test. The reason for this difference is that, unlike the student-level outcomes, the expected direction of the program's effect on teacher retention is ambiguous.

3. Methods to Maximize Response Rates and Deal with Nonresponse

We do not anticipate a high level of grantee nonresponse since evaluation grantees have demonstrated the ability and willingness to participate in the evaluation. As an evaluation grantee, districts will benefit from extensive technical assistance and receive a minimum $1 million that may be used to meet the costs associated with their participation. Our efforts, in terms of evaluation grantees, will be to ensure their successful participation in the study.

If the evaluation competition does not result in a sufficient sample of schools, we will follow the steps described above to recruit grantees from the main competition and emphasize to them the benefits of participation.

4. Pilot Testing

We will not conduct any pilot testing for the recruiting phase of the evaluation.

5. Individuals Consulted on Statistical Aspects of the Design

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, MPR	609-716-4391
Steven Glazerman	Senior Researcher, MPR	202-484-4834
Matthew Springer	Research Assistant, Professor of Public Policy and Education, and Director, National Center on Performance Incentives, Vanderbilt University	615-322-5524
John Deke	Senior Researcher, MPR	609-275-2230
Alison Wellington	Senior Researcher, MPR	202-484-4696
Dan Player	Researcher, MPR	609-945-3368
Hanley Chiang	Researcher, MPR	617-674-8374

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

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

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