U.S. Department of Education

Evaluation of the Impact of Teacher Induction Programs

Office of Management and Budget
Statement for Paperwork Reduction Act Submission

Part B: Collection of Information Employing Statistical Methods

Contract ED-04-CO-0112/0001

February 26, 2008

CONTENTS

Chapter **Page**

This package represents a request for a short extension of 9 months for data collection instruments previously approved by OMB (OMB Control No. 1850-0802, approval notice dated August 16, 2005). The clearance initially granted was for a period of 3 years, with an expiration date of August 31, 2008. Data collection for the final administration of the teacher retention survey (Appendix I) is planned to begin in October 2008, and therefore an extension on the clearance is needed. Because the design for and burden of the final round of data collection was included in the original package, this current package is identical in content to the package approved by OMB. (Minor changes in wording have been made to the section headings to reflect the current OMB headings.)"

B. COLLECTION OF INFORMATION EMPLOYING STATISTICAL METHODS

1. Respondent Universe and Sampling Methods

The study does not aim to form a statistically representative sample of a national population. Rather, our goal is to achieve a sample that includes school districts that represent a variety of policy-relevant contexts in which to observe the effectiveness of high-intensity teacher induction programs. For example, we want to exclude the few districts that already have such a program in place. We also want the districts to be geographically diverse, so that our results will be relevant for different regions of the country. Finally, we want to ensure that the districts serve disadvantaged students and are likely to have a challenge finding good teachers, so that the high-intensity induction programs have the potential to bring about positive change.

The final sample of districts will be a convenience sample. Districts are being recruited by reliance on the extensive personal networks of a subcontractor, the Center for Educational Leadership (CEL). CEL staff include former superintendents who are on good terms with current district and state education officials around the country. Relying on CEL’s networks to recruit districts is worthwhile, since it is likely to lead to much lower costs than if MPR were, in the absence of personal connections, to contact districts. It is also a reasonable approach, because the network of contacts is extensive and reaches to all regions of the country. To protect against idiosyncrasies in the sample produced by this method, we have supplemented the sample with a set of districts that meet all our criteria but are unknown to CEL staff. Given this sampling strategy, the results will be presented so that it is clear that the results are internally valid, but not representative of all districts nationwide.

Within districts, our approach is to select a set of schools to participate in the study and then randomly assign approximately half of those schools to a treatment group whose eligible beginning teachers will receive high intensity teacher induction services and half to a control group whose eligible beginning teachers will receive the usual induction services offered by the district.

2. Information Collection Procedures

Below, we describe in greater detail the rationale for our study design and the process we are using for selecting school districts, schools, and teachers for the study.

a. Statistical Methodology for Stratification and Sample Selection

In this section, we discuss four aspects of the study design and sample selection: (1) determining and achieving the target sample size of teachers, (2) selecting and recruiting school districts, (3) assigning districts to the two treatment programs, and (4) assigning schools to the treatment and control groups.

Determining and Achieving the Target Sample Size of Teachers. The fundamental unit of analysis is the teacher, so an important component of the study design was determining the number of teachers required for the study to achieve statistically precise estimates of program impacts. We have determined that the appropriate number of teachers to include in the study is 960. Assuming that there will be approximately 2.4 eligible new teachers per school, this corresponds to a sample with about 400 schools. If we spread those over 20 districts, the sample would have 20 schools per district, with 10 schools each in the treatment and control groups, or 24 teachers in each group, on average, within each district.

We arrived at this sample size requirement by setting the minimum size impact that would be meaningful to policy makers and ensuring that, if the impact were that low, that the study would be able to detect it using conventional levels of statistical significance (5 percent, for a two-sided hypothesis test) and statistical power (80 percent). This sample allows us to detect impacts on retention outcomes that are at least 5.5 percentage points and impacts on student achievement that are at least 10 percent of a standard deviation (under optimistic assumptions). These are also known as "minimum detectable impacts" (MDIs). We discuss the details of the statistical power calculations in subsection c below.

Selecting and Recruiting School Districts. Once the design was selected, we needed to define criteria for selecting school districts and develop plans to recruit them. To select districts, we used two criteria: size and poverty. Size was measured as the number of eligible elementary schools and/or eligible teachers. Choosing a threshold for district size involved balancing competing concerns. On the one hand, including only large districts ensures against a risk of having too few eligible schools in the study. In addition, the study may be easier to implement in large districts, since they are more likely to have formalized hiring processes that meet specified deadlines. On the other hand, restricting the sample to very large districts might limit the generalizability of the study's findings.

We chose to study only elementary schools for several reasons. First, a randomized trial studying teacher induction was only feasible at the elementary level. This is because it is not usually possible to vary the induction services within schools, so instead we had to have the same sample of teachers spread out over more schools. This is more easily done at the elementary level. Second, including secondary schools would unnecessarily complicate the analysis and reduce our ability to detect accurately the impacts of the high-intensity induction programs. There are important implementation issues that would differ by school level, including the selection of the mentor, the departmentalization of teachers at the secondary level, and the focus of the mentoring activities. For example, induction programs for elementary teachers would probably focus more on content-matter support, while those for secondary school teachers would focus more on pedagogical support. In addition, receptivity to the study is likely to differ by level, since there exists a perception that secondary schools historically are more resistant to change. Finally, the labor market opportunities for teachers at these two levels may differ—which means that principals of elementary schools and those of high schools would face different challenges in recruiting and retaining teachers. The effects of teacher induction at the middle and high school levels could be studied in future research.

The second district selection criterion, the concentration of poverty, was measured by setting a threshold percentage of students in each school who are eligible for free or reduced-price lunch. Districts with a concentration of schools that exceeded this threshold were determined most appropriate for inclusion in the study, since those districts are likely to have chronic problems with teacher shortages.

We also considered the *percentage* of the district's schools that meet the poverty criterion, since districts may be reluctant to have the study dictate which schools are to be included. If the percentage of schools in a district that meet the poverty criterion is too low, we risk creating a sample that does not meet the goal of having "high-need" schools.

To implement these criteria, we established specific cutoffs using the National Center for Education Statistics' Common Core of Data (CCD):

- The district had at least 15 elementary schools that qualified for Title I schoolwide assistance, which means that at least 50 percent of their students qualify for free and reduced-price lunches. A school was defined as elementary if it had at least one student in grades 1 to 4 and no students in grades 9 to 12. We required that the districts have at least 15 elementary schools, since it is likely that this cutoff would allow us to obtain an average of 20 schools per district.

- The district had at least 571 teachers in elementary schools that are eligible for Title I schoolwide assistance.[1] An eligible teacher is a regular classroom full-time equivalent in an eligible school.
- At least 70 percent of the district's elementary schools qualify for Title I assistance.

Assigning Districts to the Two Treatment Programs. The design calls for ETS to implement its high-intensity teacher induction program in one half of the districts and for NTC to implement its program in the other half. Our plan assigns districts to programs at random, with some restrictions imposed on the random assignment. First, we will make deterministic assignments for those districts and states where one of the two models (that of ETS or that of NTC) is already on schedule to be implemented in the future. Second, we will use district size (measured by the number of expected eligible teachers per district) as a stratifier. This will be done to ensure that the sample size is maximized for each of the two providers. While random assignment will be used, the number of districts is very low relative to the likely variation in district characteristics, such as the nature of the low-intensity induction program in place. Therefore, we do not intend to make direct comparisons between the ETS and NTC models of teacher induction programs.

Randomly Assigning Schools to the Treatment and Control Groups. Because some districts may have substantially more schools than we want in the study, we will first need to sub-sample schools within those districts. To do this, we will identify schools that are eligible for Title I schoolwide assistance and select a random sample of those to include in the study. If districts want to include or exclude certain schools in the study as a condition of participating,

[1] Requiring at least 571 eligible teachers is the equivalent of the 15-elementary-school rule, if there are 2.4 novice teachers per school and 6.3 percent of all teachers are novices, since 15 schools × (2.4 novice teachers/school) × (1 teacher ÷ 0.063 novice teachers) = 571 teachers.

however, we will conduct random assignment from among the subset of volunteer schools and draw inferences for the results based on the characteristics of the schools in the sample.

Random assignment of schools to treatment conditions is fairly straightforward, although we do intend to impose some constraints. Specifically, we will use stratification methods to ensure as even a mix as possible of schools whose teachers are in the same grade levels. That is, we do not wish to have a dramatic imbalance, for example, where the treatment group largely consists of fifth grade teachers and the control group largely first grade teachers. To the extent possible, we will also use stratification to ensure balance according to other characteristics, such as number of teachers and student demographics.

b. Estimation Procedures

The plans for the statistical analyses of the data, including descriptive statistics and multivariate models, are presented in Section A.16. To summarize, the main analysis will estimate the relationship between assignment to treatment status (either a high-intensity induction program or the low-intensity induction program normally operated by the districts) and outcomes of interest, such as teacher mobility, teacher practices, and student outcomes.

c. Sample Size Requirements

As explained in subsection (a) above, we used precision standards derived from other evaluations and nonexperimental research on teacher induction to determine that meaningful impacts can be detected through the use of a design that includes about 960 teachers. Table 4 displays MDIs for teacher retention outcomes measured in percentage points for two-tailed hypothesis tests with 80 percent power and using a 5 percent significance level.

The study will need a sample size that is large enough so that if there is an impact, we can detect it, meaning we can distinguish it from chance differences that arise from sampling variation. We estimated the MDI for several outcomes under a variety of different assumptions and determined that the optimal sample size would be 960 teachers. We assume these teachers would be distributed across roughly 400 schools, or 2.4 eligible beginning teachers per school, and evenly distributed between treatment and control groups within approximately 20 school districts.

A sample of this size will allow us to detect an impact on teacher mobility outcomes, which are expressed as percent with a move, of about 7 percentage points; an impact on student achievement after the first year of about 0.10 to 0.12 of a standard deviation; and an impact on teacher practices of about 0.22 to 0.25 of a standard deviation. For subgroup analysis, the MDIs will be larger. We intend to examine impacts by subgroups, such as induction provider type or district size, that are broken into groups that are usually no smaller than 1/3 of the sample. The assumptions that underlie our calculations and the MDIs associated with each set of assumptions are shown in Tables 4 and 5.

The rationale for achieving MDIs of this size has to do with the expected size of the impacts and the minimum size of an impact to be policy relevant. For mobility outcomes, past nonexperimental research suggests that we might expect to see impacts on retention after one or two years to be in the range of 5 to 20 percentage points. For student achievement outcomes, we believe that the impacts are unlikely to be large, so we have set the MDI to a level (0.10) that represents the smallest threshold below which we think an impact would not be educationally meaningful. Many proven education interventions have impacts that range from 0.15 to 0.80 of a standard deviation. In terms of classroom practices, we also expect impacts to be relatively small after one year. While the MDI cannot be set as low as for student achievement outcomes, we will be able to detect meaningful impacts on practice (at a level of 0.22).

TABLE 4

MINIMUM DETECTABLE IMPACT (MDI) ON TEACHER RETENTION UNDER ALTERNATIVE ASSUMPTIONS

Assumed Turnover Rate in the Absence of Intervention	Predicted Retention Rate (Percentage Points)		
	Control	Treatment	MDI
10%	90%	96%	5.5%
15%	85%	92%	6.5%
20%	80%	87%	7.3%
25%	75%	83%	7.9%
30%	70%	78%	8.3%

Note: Additional Assumptions:
Intraclass correlation = 0.10
R^2 = 0.20
Study attrition rate = 10%
Significance level = 5% (two-sided test)
Power = 80%

TABLE 5

MINIMUM DETECTABLE IMPACT (MDI) ON STUDENT ACHIEVEMENT UNDER ALTERNATIVE ASSUMPTIONS

Assumption	ICC_1	ICC_2	R^2	Teachers	Schools	MDI (Effect Size)
Availability of Pretest						
Post-test and pretest	0.10	0.10	0.50	960	400	0.10
Post-test only	0.10	0.10	0.10	960	400	0.11
Intra-Class Correlations						
Medium	0.15	0.15	0.10	960	400	0.13
High	0.20	0.15	0.10	960	400	0.14
Unavailable Test Scores (Grade Levels)						
1/5 of teachers	0.10	0.10	0.10	768	360	0.12
2/5 of teachers	0.10	0.10	0.10	576	320	0.14
3/5 of teachers	0.10	0.10	0.10	384	280	0.19
Unavailable Test Scores (Districts and Grades)						
1/5 of districts and no extra teachers	0.10	0.10	0.10	768	320	0.12
1/5 of districts and 1/5 of teachers	0.10	0.10	0.10	614	288	0.14
1/5 of districts and 2/5 of teachers	0.10	0.10	0.10	461	256	0.16
1/5 of districts and 3/5 of teachers	0.10	0.10	0.10	307	224	0.22

Note: ICC_1 is the intraclass correlation coefficient for schools.
ICC_2 is the intraclass correlation coefficient for teachers..
R^2 is the fraction of variance in test scores explained by classroom level covariates.

d. Unusual Problems Requiring Specialized Sampling Procedures

We do not anticipate any unusual problems that require specialized sampling procedures.

e. Use of Periodic Data Collection Cycles to Reduce Burden

The survey data collection activities include one mentor background survey in August 2005, one baseline teacher survey in October 2005, three teacher induction activity surveys in the 2005-2006 school year, and three retention surveys—one each during the 2006-2007, 2007-2008, and 2008-2009 school years. The mentor survey is estimated to take only 10 minutes and will be administered when mentors are gathered for training. So that burden on teachers is reduced, the first teacher induction survey will be conducted at the same time as the baseline survey. Since induction activities will change over the course of the school year, it is important to conduct three induction surveys to minimize potential problems with recall bias.

Non-survey-based data collection will be minimally burdensome. The observations of teachers' classes will be conducted in spring 2006, during two consecutive school days. Observing each teacher's classroom twice instead of only once will allow us to obtain a richer perspective on the teacher's practices, but scheduling the observations consecutively will reduce burden due to logistical issues. The collection of teachers' SAT or ACT scores and of classroom records will occur only once for each teacher.

3. Methods to Maximize Response Rates

If teachers who do not respond to surveys are substantially different from those who do, then the impact estimates could be biased. However, we think the potential problems associated with nonresponse will be minimal, because we expect to achieve high response rates for all surveys. We anticipate a 100 percent response rate for the baseline mentor and teacher surveys and the three teacher induction activities surveys; we expect to achieve this rate since these surveys will be conducted during the 2005-2006 school year and since mobility rates are very low during a school year. Therefore, for these surveys, nonresponse is not likely to be a concern. For the surveys on teacher retention, we anticipate achieving a 97 percent response rate in the 2006-2007 academic year and a 94 percent response rate in the following two years (2007-2008 and 2008-2009).

For all surveys, several steps will be taken to maximize response among sampled teachers. The surveys will be mailed directly to teachers at their schools, either their original schools or any schools to which they may have moved. MPR staff will follow up with nonrespondents and administer the survey over the telephone at the teacher's convenience. Initially, our contact information will be obtained from the information that respondents provide on the baseline teacher survey. If those contacts are unsuccessful, we will search major national locator databases, such as LexisNexis and Accurint, in an attempt to obtain additional information on the participants. If the telephone locating efforts are unsuccessful, we will dispatch trained field locaters from our national pool to conduct in-person locating for missing sample members.

Our predicted response rates are ambitious. If response rates to follow-up surveys fall below our targets, or if there was differential nonresponse in data collection on the study's outcomes, we will make statistical adjustments for impact estimates to be representative of the full sample. We will examine the extent of nonresponse bias by comparing the baseline characteristics of respondents and nonrespondents. We will also compare the characteristics of respondents in the treatment and control groups. We will conduct statistical tests (t-tests and chi-squared tests) to gauge whether the differences in characteristics of these groups are statistically significant. The methods described here can be used to form nonresponse adjustments if one or more schools do not provide student records data, or if classroom observations cannot be completed, or if those observations are determined to be unreliable for some reason.

Accounting for nonresponse will involve two approaches. We will use regression models to adjust for differences in the observable baseline characteristics of respondents in the treatment and control groups. We also will construct nonresponse weights that weight respondents according to their similarity to nonrespondents. The more similar a respondent is to nonrespondents, the more heavily that respondent will be weighted in our analyses.

These weights will be constructed by using baseline characteristics to predict response at followup. Specifically, we will run a logistic regression of follow-up response status on baseline variables. Using the parameter estimates from this regression, we will calculate the predicted probability of responding at followup for every member of the baseline sample. The inverse of these predicted probabilities will be the nonresponse weights. Finally, we will explore the sensitivity of our impact estimates to nonresponse by calculating impacts with and without the nonresponse weights.

4. Tests of Procedures

Developing the data collection forms involved preparing three teacher surveys: the baseline teacher survey, the induction activities survey, and the teacher retention survey. We designed all surveys for both interviewer and self-administration, and each was subjected to a cognitive pretest with up to nine respondents. The pretest sample was made up of teachers similar to those who will participate in this project. Careful pretesting provides a quality review on instrument wording, skip logic, transitions, and response burden to participants. With the cognitive pretest methodology, we also monitored and debriefed respondents to assess respondent comprehension, clarity of instruction, question flow, and organization. The mentor questionnaire was designed for self-administration only, as the mentors will complete the survey during the summer of 2005 training sessions conducted by NTC and ETS. The pretest survey questionnaire lengths provided the estimate of respondent burden for each instrument.

5. Individuals Consulted on Statistical Aspects of Design

The following people were consulted on statistical aspects of the study design:

- Roberto Agodini, Mathematica Policy Research, Inc., 609-936-2712
- John Deke, Mathematica Policy Research, Inc., 609-275-2230
- Mark Dynarski, Mathematica Policy Research, Inc., 609-275-2397
- Steven Glazerman, Mathematica Policy Research, Inc., 202-484-4834
- John Hall, Mathematica Policy Research, Inc., 609-275-2357
- Amy Johnson, Mathematica Policy Research, Inc., 609-936-2714
- Neil Seftor, Mathematica Policy Research, Inc. 202-484-4527
- Sarah Senesky, Mathematica Policy Research, Inc. 609-275-2365
- Thomas Smith, Vanderbilt University, 615-322-5519

This group consists of people with extensive experience in the design and analysis of randomized social experiments. One staff person is a sampling statistician, while others are labor economists, econometricians, and other methodologists.

REFERENCES

Alliance for Excellent Education. “Tapping the Potential. Retaining and Developing High-Quality New Teachers.” Washington, DC: Alliance for Excellent Education, 2004.

Benner, A.D. *The Cost of Teacher Turnover*. Austin, TX: Texas Center for Educational Research, 2000.

Hanushek, Eric A. “Some Simple Analytics of School Quality.” National Bureau of Economic Research Working Paper no. 10229. Cambridge, MA: NBER, 2004.

Hanushek, Eric A., John F. Kain, and Steven G. Rivkin. “Teachers, Schools, and Academic Achievement.” National Bureau of Economic Research Working Paper no. 6691. Cambridge, MA: NBER, 1998.

Ingersoll, R.M. “Is There Really a Teacher Shortage?” Philadelphia, PA: University of Pennsylvania, Center for the Study of Teaching and Policy and the Consortium for Policy Research in Education, 2003.

Mayer, Daniel, John Mullens, and Mary Moore. “Monitoring School Quality: An Indicators Report.” Report prepared for the U.S. Department of Education, National Center for Education Statistics. NCES 2001-030. Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, December 2000.

Sanders, W.L., and J.C. Rivers. “Research Progress Report: Cumulative and Residual Effects of Teachers on Future Student Academic Achievement.” Knoxville, TN: University of Tennessee Value-Added Research and Assessment Center, 1996.

Smith T.M., and R.M. Ingersoll. “What Are the Effects of Induction and Mentoring on Beginning Teacher Turnover?” American Educational Research Journal, vol. 41, no. 2, summer 2004.

Smith, T.M., and R.M. Ingersoll. “Reducing Teacher Turnover: What Are the Components of Effective Induction?” Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL, 2003.

Wenglinsky, Harold. “The Link Between Teacher Classroom Practices and Student Academic Performance.” *Education Policy Analysis Archives*, vol. 10, no. 12, February 2002.