AMERICAN INSTITUTES FOR RESEARCH®

Impact Evaluation of
Teacher and Leader Evaluation Systems

OMB Clearance Request

For Data Collection Instruments, Part B

October 19, 2012

Prepared for:
U.S. Department of Education
Contract No. ED-IES-11-C-0066

Prepared by:
American Institutes for Research

Contents

Exhibits

Appendixes

B. Description of Statistical Methods

The proposed study focuses on the implementation and the impacts of a package of teacher and leader evaluation system components that is consistent with current federal policy. To conduct the study, we randomly assigned 10–24 schools to 2 groups in each of the 9 participating districts: a control group will continue using the district's current teacher and leader evaluation system; a treatment group will implement the study's package of teacher and leader evaluation system components, which includes feedback on instructional practice, principal leadership, and student growth. We will collect outcome data from both groups. Due to random assignment, the average outcome levels in the control schools not receiving the treatment represent a reliable estimate of the outcome levels that would have been observed in the treatment schools had they not received the treatment. Therefore, the difference in the average outcomes between the treatment schools and the control schools within the same district represent a reliable estimate of the treatment's impact.

This approach to impact analysis is known as the "intent-to-treat" approach, in which all members of the treatment and the control groups are included in the impact analysis regardless of their actual participation in the treatment. Following this approach, we will assess the impacts of the evaluation system components on student achievement by comparing the treatment and the control schools in average reading and mathematics achievement, regardless of the extent to which teachers at each school actually participated in the teacher and leader evaluation system activities associated with the treatment. The impacts of the treatment will be estimated separately for each district and then pooled across districts to create an average impact of the treatments (as in a meta-analysis). The resulting intent-to-treat estimates can be interpreted as the impact of being assigned to implement the study's teacher and leader evaluation system, rather than the impact of participating in those activities. In some respects, these estimates mirror those likely to be observed in real-world settings.

In the remainder of Part B, we address the following: respondent universe and sampling, procedures for data collection, procedures to maximize response rates, pilot-testing the instruments, and the names of statistical and methodological consultants and data collectors.

1. Respondent Universe and Sampling Methods

AIR established the sample of participating districts with a multistep process approved by OMB as described in the first OMB submission for the TLES Study (OMB 1850-0890). Because the TLES Study does not employ random sampling of districts or schools for the sake of generalizability, these districts were screened and recruited on the basis of the characteristics required by the study design.

In the first step, AIR analyzed extant data on state policy. All districts in 21 states were deemed ineligible due to state initiatives in teacher and leader evaluation that would eliminate or reduce the service contrast in either 2012–13 or 2013–14.

Within the remaining 29 states, the Common Core Data from the U.S. Department of Education (2010) was then used to identify 457 districts that were of sufficient size. For a district to be eligible for the study, it was required to have at least 10 schools, with at least 6 being elementary schools.

Of the 457 districts, 100 expressed interest in speaking with us. For each of these districts, the study team interviewed the district contact as well as a district data/assessment expert via telephone using the screening protocol approved by OMB in May 2012. The screening protocol was used to determine the district's eligibility to participate in the study, based on its intended evaluation system practices for 2012–13 and 2013–14 and the adequacy of its data systems for value-added modeling. The following criteria were used in this determination:

- **Teacher evaluation system.**

  - The current classroom observation protocol used for evaluating teachers is not comparable to FFT or CLASS or is not implemented intensively (e.g., does not include comprehensive yearly training from the provider of the protocol).

  - Tenured teachers are observed at most twice a year. Teachers on probation or teachers who are identified as having performance issues may be observed more often.

  - The current teacher evaluation system does not include a teacher effectiveness measure based on value-added modeling.

- **Leader evaluation system.** The current leader evaluation system is not using a 360-degree assessment tool.

- **Data system.** The current data systems include student assessment data and teacher-student-course ID linkages required for conducting value-added modeling.

The screening process and subsequent recruitment conversations with interested districts resulted in 19 site visits to eligible and interested districts. Nine districts agreed to participate. Within these districts, eligible schools were identifed through dialogue with district officials about competing intiatives and other barriers to school participation, and presentations about the study were made to the principals of the eligible schools. A total of 140 schools in the 9 districts signed memoranda of understanding indicating their willingness to be randomly assigned as part of the study.

The final sample includes 4 districts using FFT and 5 districts using CLASS. Within each participating district, the participating schools have been randomly assigned to one of two groups: (1) piloting the evaluation system components provided through TLES (i.e., the treatment group) and (2) continuing with "business as usual" (i.e., the control group).

We anticipate approximately 21 mathematics and reading teachers per school, each teaching an average of 25 students, in a given academic year. Thus, the total universe of teachers will be about 2,940; the total universe of students will be about 73,500. (See Exhibit 1 for the complete structure of the sampling design.)

Exhibit 1. Sampling Design

District	Teacher Observation Instrument	Study Group	Number of Schools (unit of randomization)	Number of Teachers (based on estimate of 21 teachers per school)	Number of Students in study (based on estimate of 25 students per teacher)
District 1	FFT	Treatment	5	105	2,625
District 1	FFT	Control	6	126	3,150
District 2	FFT	Treatment	8	168	4,200
District 2	FFT	Control	7	147	3,675
District 3	FFT	Treatment	12	252	6,300
District 3	FFT	Control	12	252	6,300
District 4	FFT	Treatment	7	147	3,675
District 4	FFT	Control	7	147	3,675
District 5	CLASS	Treatment	9	189	4,725
District 5	CLASS	Control	9	189	4,725
District 6	CLASS	Treatment	7	147	3,675
District 6	CLASS	Control	6	126	3,150
District 7	CLASS	Treatment	11	231	5,775
District 7	CLASS	Control	11	231	5,775
District 8	CLASS	Treatment	6	126	3,150
District 8	CLASS	Control	7	147	3,675
District 9	CLASS	Treatment	5	105	2,625
District 9	CLASS	Control	5	105	2,625
Totals			**140**	*2,940*	*73,500*

To assess the statistical power of the study design, we draw on recent literature on power analysis for group randomized trials (Schochet, 2008; Spybrook, Raudenbush, Congdon, & Martinez, 2009) to calculate the variance components and estimate the minimum detectable effect sizes (MDESs) for student achievement outcomes, teacher practice outcomes, teacher mobility outcomes, and intermediate outcomes (i.e., decisions of key actors) as measured by the teacher survey. We derive assumptions from prior studies about the proportion of the variance in the outcome measures that are between schools and between teachers within schools, the percentage of outcome variance explained by covariates, the number of districts and the number of schools per district, the number of teachers per school, the number of students per teacher, and the number of teachers observed per school. To reflect both optimistic and cautious assumptions, we have calculated MDES ranges for our main outcome measures (Exhibit 2).

Exhibit 2. MDES for Main Outcome Measures

Outcome Measure	MDES (Optimistic—Cautious)
Student achievement	0.08–0.11
Teacher practice	0.15–0.16
Teacher mobility	9.4–9.7 percent (based on 20% base mobility rate in control group)
Intermediate outcomes as measured by the teacher survey	0.18

2. Procedures for Data Collection

AIR project staff will manage data collection and ensure quality and timeliness. The data collection instruments for which clearance is requested in this submission are included in Appendixes A–D. They include the teacher survey, the principal survey, the district interview, and the district archival records collection protocol, as summarized in the study description preceding Part A of this submission. The teacher survey will be administered online to all teachers responsible for reading or mathematics instruction in any of Grades K–8. The principal survey will be administered online to all the principals of the study schools. The district interview will be conducted via telephone with each study district. Archival record requests will be sent via e-mail to each study district. The above data collections are specified in the following timeline:

1. **January 2013.** Student records, and teacher and principal records for online survey administration.

2. **April–May, 2013.** Teacher and principals surveys, and district interviews.

3. **July–September, 2013.** Student records, local teacher performance evaluation ratings, and teacher and principal records for mobility analyses.

4. **January–March, 2014.** Teacher and principal records for online survey administration.

5. **April–May, 2014.** Teacher and principals surveys, and district interviews.

6. **July–September, 2014.** Student records, local teacher performance evaluation ratings, and teacher and principal records for mobility analyses.

3. Procedures to Maximize Response Rates

Based on our extensive experience with administering surveys in a variety of schools, districts, and states, including a recent Intensive Partnerships for Effective Teaching (IPS) study funded by the Bill & Melinda Gates Foundation, we anticipate the response rate to be approximately 85 percent for the teacher and the principal surveys. We anticipate a 100 percent response rate for district interviews and the archival records requests. We reference the IPS study in particular because it is the most recent example of teacher and principal surveys conducted by AIR around the issue of evaluation systems. The IPS study achieved response rates of 81 percent on the teacher survey and 76 percent on the school leader survey. However, because the IPS surveys were longer than our proposed surveys by roughly 60 percent, we believe that our response rate approximations are appropriate estimates.

The following procedures will be used to ensure high response rates:

- Obtaining high response rates depends in part on the quality of the instruments. The team will pilot and subsequently refine all instruments to ensure that they are user-friendly and easily understandable, which will increase participants' willingness to participate in the data collection activities and thus increase response rates. See the next section for information on procedures designed to ensure instrument quality.

- Obtaining high response rates also depends in part on the length of the instruments. The teacher and principal surveys require an administration time of approximately 30 minutes. The district interview is restricted to 90 minutes, which is reasonable given that districts are highly motivated to participate in the study.

- To further ensure a high response rate on the teacher survey, AIR will not rely entirely on Web-based administration. AIR will conduct follow-up activities with telephone prompts, as necessary, and a hard-copy pencil-and-paper survey questionnaire will be mailed to any respondent who requests one. Approximately 2 weeks after the initial mailing, we will begin the process of survey follow-up. We will send a letter reminding respondents about the survey. After 2 more weeks, we will implement a series of 3 follow-up calls at approximately 10-day intervals. During the third call, we will offer to complete the questionnaire as a telephone interview. Using these procedures, the research team has extensive experience administering Web- and e-mail-based surveys with high response rates.

- District coordinators employed by the study will be responsible for maintaining contact with the respondents as well as garnering the support of school principals for encouraging survey completion.

- The study will offer a social incentive to the respondents by stressing the importance of the data collections as part of a high-profile study that will provide much-needed information to the districts and the schools.

- Teacher survey respondents in both the treatment and the control groups will receive a small amount of compensation in return for participating in the data collection activities. This is to make them feel that we value their time and participation, thus encouraging them to participate and increasing the response rate. For specific details, please see Part A section 9, Payment or Gifts.

4. Pilot-Testing Instruments

The teacher and principal surveys will be pilot-tested with small numbers of respondents (fewer than 10 respondents per instrument) and revised to ensure that the questions are as clear and simple as possible for the respondents to complete.

Pilot test subjects will include teachers and principals who are in similar situations as the study's treatment educators (e.g., participating in a pilot system while an existing evaluation system is in place. A think-aloud, or cognitive lab, format will be used for pilot testing, whereby the respondents will be asked to complete the draft instrument, explain their thinking as they constructed their responses, and identify the following:

- Questions or response options that are difficult to understand.

- Questions in which none of the response options is an accurate description of a respondent's circumstance.

- Questions that call for a single response but more than one of the options is an appropriate response.

- Terms that are not defined that should be defined.

- Questions for which the information requested is unavailable.

5. Names of Statistical and Methodological Consultants and Data Collectors

This project is being conducted by AIR under contract to the U.S. Department of Education. Michael Garet is the principal investigator, and Andrew Wayne is the project director. The senior task leaders from AIR contributing to the study methods and data collection are Seth Brown, Jinok Kim, Anja Kurki, and David Manzeske. The instruments were developed by Michael Garet, Andrew Wayne, David Manzeske, Seth Brown, and additional project staff at AIR. AIR project staff will carry out the data collection activities.

References

Schochet, P. (2008). Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics, 33*(1), 62–87.

Spybrook, J., Raudenbush, S. W., Congdon, R., & Martinez, A. (2009). *Optimal design for longitudinal and multilevel research: Documentation for the Optimal Design software.* Ann Arbor: University of Michigan.