AMERICAN INSTITUTES FOR RESEARCH®

Impact Evaluation of Teacher and Leader Evaluation Systems

OMB Clearance Request, Part A

February 17, 2012

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

Prepared by:
American Institutes for Research

Contents

List of Exhibits

Introduction

A core premise behind federal policies related to teacher quality—the 2002 reauthorization of the Elementary and Secondary Education Act (ESEA), the American Reinvestment and Recovery Act (ARRA) of 2009, and the U.S. Department of Education's (ED's) 2010 Blueprint for the reauthorization of ESEA—is that teacher quality is a potential lever for improving student achievement. Several studies support this premise, with recent work examining individual teachers over multiple years to estimate the degree of variation in achievement gains that can be attributed to teachers rather than measurement error (see Bill & Melinda Gates Foundation, 2011; Schochet & Chiang, 2010). Similar studies suggest that some variation in achievement gains can be attributed to principals, which indicates that principal quality may be another lever for policymakers (Hallinger & Heck, 1998; Leithwood, Harris, & Strauss, 2010; Leithwood & Jantzi, 2005; Leithwood, Louis, Anderson, & Wahlstrom, 2004; Waters, Marzano, & McNulty, 2003).

Systems for measuring teacher performance traditionally have relied on checklist measures that capture several aspects of instruction and a small number of other criteria (e.g., parent relations). Principals or other school administrators are expected to observe teachers and complete the checklist as part of an evaluation cycle, with the number of observations depending on the teacher's seniority. Many concerns have been expressed about the validity of this approach, including, for example, the lack of specificity in checklist instruments, lack of principal training to use the instrument, and lack of time for principals to fulfill these responsibilities (Danielson & McGreal, 2000; Porter, Youngs, & Odden 2001; Weiss & Weiss, 1998). As implemented, these systems have reportedly failed to identify differences in teacher performance. In a recent study of teacher performance measurement practices in 10 districts, researchers found that more than 99 percent of teachers were rated as satisfactory (Weisberg, Sexton, Mulhern, & Keeling, 2009).

Through competitive grant programs—specifically the Teacher Incentive Fund (TIF) and Race to the Top—the federal government has been supporting state and district efforts to improve the evaluation of teacher and leader performance and to use that data in human resource policies. Both programs encourage the use of student achievement growth as a performance measure for both teachers and leaders. At the same time, additional performance measures are encouraged. For teachers, one of the main additional measures is classroom practices. For leaders, federal program guidance calls for the use of student growth information as well as other measures that produce data that are reliable and transparent, differentiate performance, and inform professional development, among other features.

Despite the potential of teacher and leader quality to affect student achievement and the prominence given to teacher and leader evaluation systems in recent policy, there are no scalable approaches to improving teacher and leader quality that have been proven to be effective through randomized controlled trials (RCTs). This study, the *Impact Evaluation of Teacher and Leader Evaluation Systems (TLES)*, will begin to fill this research gap. The purpose of this study is to examine the implementation of a teacher and leader evaluation system that has the features called for in federal policy and its impacts (e.g., impacts on teacher mobility, impacts on student achievement). To this end, the study will involve the recruitment of approximately 14 districts and 182 schools and collection of archival data as well as teacher and administrator surveys, teacher observations, and telephone interviews with one official from each participating district. This study is funded by the Teacher Incentive Fund, as authorized by the Departments of Labor, Health and Human Services, and Education, and Related Agencies Appropriations Act, 2010, Division D, Title III, Pub. L. 111-117 (2010 Consolidated Appropriations Act, 2011, Pub. L. 112-10).

The Institute of Education Sciences (IES) of the U.S. Department of Education (ED) requests clearance for the study's district-level screening protocol to be used for recruitment. Subsequent to the approval of the current package, a second package will be submitted to request clearance for the data collection instruments.

This package contains two major sections with multiple subsections:

- Description of the Impact Evaluation of Teacher and Leader Evaluation Systems
  - Purpose
  - Research Questions
  - Treatment Selection and Characteristics
  - Experimental Design, Analytic Strategy, and Sample
  - Data Collection
- Supporting Statement for Paperwork Reduction Act Submission
  - Justification (Part A)
  - Description of Statistical Methods (Part B)

Description of the Impact Evaluation of Teacher and Leader Evaluation Systems

Purpose

The TLES study is designed to examine the implementation and impact of a teacher and leader evaluation system that is consistent with current federal policy. In contrast to traditional systems, the system provided by the study will use multiple, valid measures, including student growth, to meaningfully differentiate performance levels. The system will provide summative performance ratings annually, drawing on multiple assessments of teacher practice and principal leadership that provide timely feedback to guide the efforts of educators and those who supervise and support them.

To determine the impacts of the study's system on student achievement and other outcomes, a purposive sample of districts with traditional evaluation systems will be recruited during the 2011–2012 school year. These districts will be selected on the basis of data infrastructure for measuring student growth (i.e., linked student and teacher records) and genuine interest in use of educator performance information to improve student achievement. A pool of eligible schools will be identified in each district and randomly assigned to treatment and control conditions, with equal numbers of treatment and control schools in each district. In the treatment schools, the study's teacher and leader evaluation system will be introduced during the summer of 2012 and implemented during the 2012–2013 and 2013–2014 school years, and school districts will be expected to support use of the performance information in those schools. In the control schools, the district's current evaluation system practices will continue to be used (i.e., business-as-usual). During the two-year implementation period, data will be collected to support analyses of the implementation and impact of the pilot evaluation system, with a final collection of data for impact analyses in fall 2014.

Research Questions

The study has six research questions designed to assess the implementation of the evaluation system provided by the study and the impacts of this system on teacher practice, principal leadership, and student achievement.

> ***RQ1. How was the intervention implemented (e.g., fidelity with which the intervention was delivered, participation of key actors)?***
>
> ***RQ2. Did the intervention produce a contrast between the treatment and control schools in teachers' and principals' experiences with evaluation of their performance (e.g., how frequently teachers reported being observed)?***
>
> ***RQ3. What impacts did the evaluation system have on the decisions of key actors (e.g., teachers' decisions to try new techniques or to pursue learning experiences)?***
>
> ***RQ4. What impacts did the evaluation system have on the mobility of low-value-added teachers and high-value-added teachers?***

RQ5. What impacts did the evaluation system have on the dimensions of teacher instructional practice and principal leadership that are the focus of the intervention?

RQ6. What impacts did the evaluation system have on student achievement?

Because the study intervention will continue for two years, we will address the descriptive question (RQ1) in reference to the first year and the second year separately. We will examine impact questions (RQ2, RQ3, RQ4, RQ5, and RQ6) separately at the end of the first and the second years.[1] Impact analyses using data pertaining to the end of the second year will measure the cumulative impact of two years of the intervention.

Intervention Selection and Characteristics

The study's evaluation system will consist of a teacher evaluation system and a leader evaluation system. Both systems will use measures of student growth (annually, using value added methods). In addition, the teacher evaluation system will measure teachers' instructional practice (two to four observations per teacher per year), and the leader evaluation system will measure principal leadership (two assessments per principal per year). Each year, these measures will be combined into a summative measure for each teacher and principal.

On the basis of these measures, the evaluation system is intended to define performance expectations, repeatedly measure performance, and produce actionable reports. In addition, the study districts are expected to adopt and implement approaches to supporting the use of the performance information. In the text below, we elaborate on each of these objectives.

Performance expectations

To define performance expectations, our study treatment will use objective, transparent measures for teacher practice, principal leadership, and student growth. Teacher practice will be measured using observation protocols grounded in frameworks for effective instruction and supported by video libraries that illustrate superior performance. Principal leadership will be measured using a survey-based assessment focused on aspects of leadership that are relevant to instructional leadership, as well as a supplemental component measuring the fulfillment of the principal's responsibilities in executing the study's teacher evaluation system. Student growth will be measured using value-added modeling of achievement data in reading and mathematics, with teachers in untested grades and subjects receiving the average value-added score. Each district will work with the study team to set district-specific performance targets for each measure as well as rules by which the separate measures will be combined to determine the composite performance labels for teachers and leaders (e.g., "highly effective," "effective," or "requiring improvement"). Level definitions, performance targets, and the methods for composite score calculations will be shared with stakeholders at the start of implementation, thus ensuring that the system's expectations will be fully transparent to teachers and principals.

[1] Because observations of instructional practice will be conducted only during the second year of implementation, impact on instructional practice [within RQ5] will be examined only at the end of the second year.

Repeated measurement of performance

Repeated measurement provides a basis for both formative and summative uses of the performance information. The study's evaluation system will repeatedly measure teachers' instructional practice (two to four observations per teacher per year), principal leadership (two assessments per principal per year), and student growth (annually, using value added methods). To support summative uses, composite scores for teachers and principals (based on teacher practice or leadership practice scores combined with student growth measures) will be calculated at the end of each year and translated into a composite performance label.

Actionable performance reports

In-person feedback and online reports will provide performance reports tailored to the needs of the educators being evaluated (i.e., "evaluees"), their supervisors, and others designated by the central office who are expected to use the performance information. For evaluees, performance reports will indicate the individual performance level in the context of clear performance expectations and accountability, to guide their efforts to perform and encourage skill development. Each measurement instance will include opportunities to discuss the measurement results. Observers and assessors will conduct these discussions using well specified protocols. These discussions will focus on the evaluee's performance and the performance expectations (Danielson, 2010; Goe, Bigger, & Croft, 2011). For supervisors and others designated by the central office, additional reporting formats will be available to illuminate patterns and support the use of summative information.

District support for the use of the performance information

To promote uses of performance information that may improve student achievement, each district will be asked to make a set of key, district-level, decisions about how the evaluation system will be used in the treatment schools in their district, and how this use will be supported. These decisions will occur during the study's recruitment phase. For example:

- Districts may choose to use the performance information to support teachers' efforts to engage in professional development opportunities that are tailored to their needs. To do so, a district might, for example, review its menu of professional development offerings to assess which activities are aligned to the study's observation frameworks, and the district may facilitate teachers' access to those offerings.
- Districts may choose to integrate the performance information into the current performance evaluation system, as was done in the schools participating in the Excellence in Teaching Program in Chicago Public Schools. In that program, principals incorporated ratings from the pilot observation measure in teacher's official evaluations by taking advantage of a policy that allows principals to supplement district criteria with "local school requirements" at the beginning of the school year.

Districts will be recruited into the study based on—among other factors—their interest in implementing evaluation system practices like those called for in federal policy.

The four intervention objectives described above will occur in the context of schoolwide implementation of the system. Including all teachers and the principal in each school is intended to increase awareness of expectations, motivation, and use of collegial support (see Garet, Porter, Desimone, Birman, & Yoon 2001).

In what follows, we describe the three main components of the planned performance measurement system: a component for feedback on student growth, a component for feedback on instructional practice, and a component for feedback on principal leadership.

1. Component for Feedback on Student Growth

At the end of each of the two intervention years, the study will generate evaluation system reports for teachers, principals, and district leaders that summarize performance as measured by student growth. Principals will receive training in their interpretation and will contribute to school-based sessions to help teachers interpret their results individually and collectively. To ensure coherence with existing school processes, principals will be encouraged to integrate these sessions into existing meetings focused on the interpretation and use of student achievement data.

2. Component for Feedback on Instructional Practice

Although a number of frameworks for instructional practice have been developed for research purposes, including several frameworks for specific content areas, two frameworks have emerged as most suitable for use in teacher evaluation systems because of their applicability across subjects and grades, and because of evidence of their validity and connection to student achievement: the Framework for Teaching (FFT) and the Classroom Assessment Scoring System (CLASS) (Goe, Bell, & Little, 2008).

The study districts will be divided between these two frameworks, such that teachers in the treatment schools in 7 districts will receive feedback on instructional practice using FFT, and those in the treatment schools in the other 7 districts will do so using CLASS. District assignment to one of the two frameworks will be nonrandom; district preferences for one or the other framework will be taken into account if doing so facilitates recruitment while preserving the balance in the number of districts assigned to each framework. Thus, the main impact findings, pooling across all 14 districts, will pertain to frameworks for feedback on instructional practice having the general features called for in recent policy developments, rather than one particular approach.

FFT and CLASS have many similarities but have different origins and theoretical underpinnings (Goe et al., 2008). As Exhibit 1 shows, the frameworks have many similar constructs and similar measurement approaches, although their rating scales are somewhat different: items are rated on a four-level scale for the FFT but a seven-level scale for the CLASS.

Exhibit 1. Domains and Constructs for FFT and CLASS

Framework for Teaching (FFT)	Classroom Assessment Scoring System (CLASS) (CLASS-Upper Elementary and CLASS-Secondary)
Domain 2: Classroom Environment • Creating an Environment of Respect and Rapport • Establishing a Culture for Learning • Managing Classroom Procedures • Managing Student Behavior • Organizing Physical Space **Domain 3: Instruction** • Communicating with Students • Using Questioning and Discussion Techniques • Engaging Students in Learning • Using Assessment in Instruction • Demonstrating Flexibility and Responsiveness	**Domain 1: Emotional Support** • Positive Climate • Negative Climate • Teacher Sensitivity • Regard for Adolescent Perspectives **Domain 2: Classroom Organization** • Behavior Management • Productivity • Instructional Learning Formats **Domain 3: Instructional Support** • Content Understanding • Analysis and Problem Solving • Quality of Feedback • Instructional Dialogue

NOTE: The FFT includes two other domains—Planning and Preparation and Professional Responsibilities—that are not amenable to measurement through observation and are not included in the study's measure of teacher practice.

The implementation of the two systems in the study districts will follow the same general parameters. For example, for teachers in grades 4-8 who are responsible for instruction in mathematics or reading/English language arts, both systems will provide four rounds of observation and feedback, and both will use peer observers for three cycles and principals for a fourth cycle. Both systems also will be similar in the extent and nature of the framework-aligned supports provided for individual development.

3. Component for Feedback on Principal Leadership

To define expectations for principal leadership, we will deploy the Vanderbilt Assessment of Leadership (VAL-ED) in treatment schools in all study districts. VAL-ED is a tool for use in principal evaluation systems with established validity and implementation fidelity (Condon & Clifford, 2009). The researchers who developed VAL-ED have published its psychometric properties in peer-reviewed journals and websites (www.valed.com/research) and have been continuously improving the tool. These factors, as well as its ease of implementation, will facilitate study recruitment. VAL-ED also is aligned with national standards for principal leadership (Goldring, Carvens, Murphy, Porter, Elliott, & Carson 2009).

VAL-ED focuses on leadership as it relates to teacher and student learning, so the teacher and leader systems will share the same broad purposes. VAL-ED gathers data on principal behaviors from principals, supervisors, and teachers (an approach labeled "360-degree assessment") and provides both criterion-referenced and norm-referenced scores. Thus, principals can understand their performance ratings in terms of specified criteria or relative to other principals evaluated using VAL-ED in the United States.

To encourage principals to fulfill their responsibilities in implementing the feedback system for teacher practice, we will complement VAL-ED with expectations for principals for compliance with the requirements of the teacher evaluation system: participation in observer trainings, conduct of observations, and conduct of debriefing sessions.

Experimental Design, Analytic Strategy, and Sample

The TLES Study will employ an experimental design with randomization of schools to treatment and control conditions within participating districts. The choice of schools as the unit of assignment is based on both theoretical and practical considerations. Theory suggests that the potential impact of well-articulated performance expectations and repeated measurement and discussion of performance can be magnified in a schoolwide implementation, both in terms of teacher and principal understanding of the performance expectations and in terms of efforts to achieve the expectations (see Garet et al. 2001). Practically, principals have the responsibility for entire schools, so randomization within school is not feasible for the principal evaluation component and would be vulnerable to contamination for the teacher evaluation component because the principal will conduct a quarter of the teacher observations.

The study team will recruit approximately 14 districts to pilot the study's teacher and leader evaluation system. Within each of the 14 districts, the study will identify approximately 12 to 14 schools and randomly assign each to either the treatment or the control condition. These schools will include both elementary schools and middle schools, so study findings will be relevant to both school levels.

An important design issue is to determine the sample size needed for estimating the impact of the intervention (RQ3, RQ4, RQ5, and RQ6) with adequate statistical power. The following sections present the analytic strategy and power analyses, first for impacts on student outcomes and then for impacts on teacher and principal outcomes.

Analytic Strategy and Statistical Power for Impact on Student Outcomes

Analytic Strategy. To estimate the treatment effect on student achievement, we propose to use the following three-level hierarchical linear model, in which students are nested within teachers/classrooms, which are in turn nested within schools.

$$Y_{ijk} = \mu + \theta T_k + \boldsymbol{\pi}' \boldsymbol{X}_{ijk} + \boldsymbol{\beta}' \boldsymbol{W}_{jk} + \boldsymbol{\gamma}' \boldsymbol{S}_k + u_k + r_{jk} + e_{ijk} \,. \qquad (1)$$

In the equation, Y_{ijk} is the reading or mathematics score on state assessment for student *i* taught by teacher *j* in school *k*. We assume that the outcome scores are standardized separately within grade level within district. Because the study randomly assigns schools to the treatment condition or the control condition, the model includes a school-level treatment indicator T_k, which is coded 1 if school *k* is assigned to the treatment condition and 0 if the school is assigned to the control condition. The coefficient θ represents the average effect of the intervention relative to the control on student performance.

The variables $\boldsymbol{X}_{ijk}$, $\boldsymbol{W}_{jk}$, and $\boldsymbol{S}_k$ are vectors of student background characteristics, teacher background characteristics, and school background characteristics, respectively, measured prior to the initiation of the intervention. The background characteristics are included in the model to control for potential remaining imbalances across study conditions and to increase the statistical precision of the impact estimates. Key student covariates include pretest scores, which are the states' annual assessment scores from the previous year.[2] Because state assessment scores for reading and mathematics are typically administered in Grades 3–8, we expect to estimate the average treatment effect on the achievement of students in Grades 4–8 using this model.[3] The residuals u_k, $r_{jk,}$ and e_{ijk} represent the random errors associated with schools, teachers, and students, respectively.

To maximize precision, we plan to pool the data across the 14 participating districts and from elementary and middle schools (Grades 4–8) in a single model. To take into account potential variation in assessments across districts and grades, we plan to include a vector of district-by-grade indicators in the model. With 14 districts and five grades (Grade 4–8) in our analytic sample, the models will thus include 70 indicators of the district-grade combinations. We also propose to include interactions between the pretest and district-by-grade indicators to take into account potential variation in the relationships between pretest and posttest scores across districts and grades. In addition, we will include interactions between treatment and district indicators to examine whether impact estimates are heterogeneous across districts.

Statistical Power. The methods used here draw on recent literature on power analysis for group randomized trials (Schochet, 2008; Spybrook, Raudenbush, Congdon, & Martinez, 2009). The power analysis determines the minimum detectable effect size (MDES) for student outcomes in student-level standard deviation units.

We assumed a two-tailed test, with 0.80 power, and a Type I error level of 0.05, as is conventional. We drew other parameters from the literature, in some cases making further refinements on the basis of analyses of archival state data sets. Below are the key parameters used in our power analyses:

- **Intraclass correlation at the school level (ICC_s).** The ICC_s is the proportion of variance in the outcome that lies between schools within districts relative to total variance. It is assumed to range from 0.07 to 0.10 as established in previous literature[4] and results from statewide data sets.[5]

[2] For the second-year student impact analysis, we will estimate the cumulative impacts of two years of implementation. The model specification will be similar to the specification estimating the first-year impacts. However, the model will control for scores in the grade they were enrolled in two years prior for the pretest variables, instead of scores in one grade prior.

[3] Because state assessments for reading and mathematics are typically administered to Grades 3-8, our primary analysis will focus on Grades 4–8 in estimating the Year 1 impact. Grade 3 is excluded because of the absence of pretest. Grades 4-8 are also the grade levels in which students will have had teachers who would receive the full intervention (i.e., feedback on student achievement growth combined with four rounds of observations and feedback).

[4] For student achievement outcomes, an ICC of 0.10 across schools within districts is in the range based on analysis of large-scale data sets (see, e.g., Bloom, Richburg-Hayes, & Black, 2007; Jacob, Zhu, & Bloom, 2009; Schochet, 2005).

[5] A statewide data set containing student achievement data from an annual state assessment yielded estimates of the unconditional ICC across schools within districts that range from 0.05 to 0.07 in various combinations of subjects and administration years.

- **Intraclass correlation at the teacher level (ICC_t).** The ICC_t is the proportion of variance in the outcome across teachers within schools relative to total variance. It is assumed to range from 0.04 to 0.07 on the basis of previous literature and results from statewide data sets.
- **Percentage of the outcome variance explained by covariates.** Baseline data on student achievement and on teacher and school characteristics would serve as covariates in the impact analyses. The percentage of outcome variance explained by the covariates is assumed to range from 50 percent to 65 percent.[6]
- **Number of districts and number of schools per district.** We assume 14 districts, half of which have approximately 12 schools per district (i.e., 6 treatment schools and 6 control schools), and half of which have approximately 14 schools per district (i.e., 7 treatment schools and 7 control schools).
- **Number of teachers per school.** For elementary schools, the analysis will focus on Grade 4 and Grade 5.[7] We assume about four teachers in each grade, or eight teachers per elementary school.[8] For middle schools, the analysis will focus only on reading and mathematics teachers, and we assume eight teachers in each subject.
- **Number of students per school.** At the elementary school level, each of the eight fourth- or fifth-grade teachers is assumed to teach 20 students, resulting in 160 students total across the two grades. The computation is more complex in middle schools because teachers typically teach several sections. If each teacher teaches approximately three classes of 20 students, it would imply 480 students. For the purpose of the power analysis, we have used the *n* of students in the elementary schools for all schools, which should provide a conservative estimate.

Exhibit 2 presents the MDES under various scenarios, ranging from more optimistic assumptions to more conservative assumptions. The MDES for student outcomes ranges from 0.07 to 0.10, which is below or equal to the MDES of 0.10 that this study expects for student outcomes.

Exhibit 2. MDES for the Impact of the Treatment on Student Achievement Under Different Assumptions for ICC and Percentage of Variance Explained by Covariates

	% variance explained by covariates	
ICC_s & ICC_t	50%	65%

[6] For student achievement outcomes, scores from previous academic year are known to explain more than 50 percent of the variability (see, e.g., Schochet, 2008). We also assume the higher percentage of 65 percent, which is close to findings from our previous studies. Data from various states, subjects, and administration years yielded 75 percent to more than 90 percent of the outcome variability explained by a district-level covariate.

[7] Because state assessment scores for reading and mathematics are typically administered in Grades 3–8, we do not expect to have the necessary prior year data on student achievement (i.e., student achievement in Grade 2) for teachers of Grade 3 in most districts.

[8] From information drawn on the Common Core of Data (the number of full-time teachers and the enrollment size by grade level) we derived the number of teachers in the elementary schools and in the middle schools.

0.10 & 0.07	0.10	0.09
0.07 & 0.04	0.08	0.07

Analytic Strategy and Statistical Power for Impact on Teacher and Principal Outcomes

Analytic Strategy. In addition to examining student achievement outcomes, the study aims to estimate the average effect of the treatment on teacher and principal outcomes. To estimate the average treatment effect on teachers, a two-level model will be used, with teachers nested within schools, whereas, to estimate the average effect on principals, a single-level model will be used. For example, to estimate the average effect of the treatment on continuous teacher outcome measures based on classroom observations, the following model can be used:

$$Y_{jk} = \mu + \theta T_{k} + \boldsymbol{\beta}' \boldsymbol{W}_{jk} + \boldsymbol{\gamma}' \boldsymbol{S}_{k} + u_{k} + r_{jk} \,. \qquad (2)$$

In the equation, Y_{jk} is a measure created from classroom observations for teacher j in school k. Given that two different observation protocols (CLASS and FFT) will be used in different districts, the teacher outcome measures will be standardized on teachers in the control schools within each district to make it possible to pool observations from all 14 districts. The model includes an indicator of the treatment schools T_k. The coefficient θ represents the average effect of the treatment on the teacher observation measure. $\boldsymbol{W}_{jk}$ and $\boldsymbol{S}_k$ are vectors of teacher and school pretreatment characteristics, respectively. The model also will include indicators of subject areas (reading or mathematics) as well as district and grade indicators. Treatment impacts on continuous measures of principal outcomes will be estimated based on school-level regressions with similar specifications.

For dichotomous teacher outcome measures based on data from classroom observations or the teacher survey, we will use hierarchical generalized linear models (HGLMs) to estimate the treatment impacts on teachers. Treatment impacts on dichotomous measures of principal outcomes will be estimated based on logistic regression.

Statistical Power. Although we propose to conduct analyses using both survey and observation measures, our power analysis focuses on the observation measures because, due to our plans for random sampling (see "Classroom Observations" subheading within the "Data Collection" section), the sample size is considerably smaller for observations. Thus, the resulting power estimate is conservative. The power analysis is based on the following assumptions:

- **Intraclass correlation at the school level ($\mathrm{ICC_s}$).** The $\mathrm{ICC_s}$ is the proportion of variance in the outcome that lies between schools within districts relative to total variance. It is assumed to be 0.02 on the basis of previous literature.[9]
- **Intraclass correlation at the district level ($\mathrm{ICC_d}$).** The $\mathrm{ICC_d}$ is the proportion of variability in the outcome across districts relative to total variance. It is assumed to be 0.02.

[9] The ICCs for attitudinal and behavioral outcomes are typically smaller than the ICCs for achievement outcomes. The ICCs for attitudinal and behavioral outcomes tend to range from 0.01 to .0.05 (Bloom et al., 2007; Jacob et al., 2009; Murray, Varnell, & Blitstein, 2004; Siddiqui, Hedeker, Flay, & Hu, 1996). Because previous literature deals with ICCs at a single level (e.g., communities, worksites, schools), it is unclear how the $\mathrm{ICC_s}$ would be divided between the school and district levels. We assumed the same level of $\mathrm{ICC_s}$ at the two levels (0.02 and 0.02).

- **Percentage of the outcome variance explained by covariates.** Baseline data on student achievement and on teacher and school characteristics would serve as covariates in the impact analyses. It is assumed that these covariates will explain 5 to 25 percent of the variance in teacher behavioral and attitudinal outcomes.[10]
- **Number of schools per district and number of districts.** We assume 14 districts, half of which have approximately 12 schools per district (i.e., 6 treatment schools and 6 control schools), and half of which have approximately 14 schools per district (i.e., 7 treatment schools and 7 control schools).
- **Number of teachers observed per school.** We assume three teachers observed per school.

Exhibit 3 presents the estimated MDES for the treatment's impact on teacher observational outcomes. The study is powered to detect an MDES of 0.24.

Exhibit 3. MDES for the Impact of the Treatment on Teacher Observation Outcomes, Under Alternative Assumptions for the Covariate R-Square

	% variance explained by covariates	
	5%	**25%**
Treatment impact	0.24	0.24

Data Collection

During this stage of the clearance process, clearance is being sought only for the district-level screening protocol used for recruitment. The screening protocol, and the procedures for the use of the protocol, are explained in detail in Part B.1 (Respondent Universe and Sampling Methods) and Part B.2 (Procedures for Data Collection). In the remainder of this section, we present the plan for data collections that will occur after recruitment is complete.

The data collection plan for the study includes collections of new data as well as collections of archival data, as shown in Exhibit 4. A description of the proposed data collection instruments follows.

[10] For teacher outcomes that are attitudinal and behavioral measures, the predictive power of covariates for analyses of teacher outcomes is likely to be weak, and covariates may or may not be available (Jacob et al., 2009).

Exhibit 4. Data Collection Plan

Data Source	Sample Size		Collection Schedule				
	Treatment Condition	Control Condition	Fall 2012	Spring 2013	Fall 2013	Spring 2014	Fall 2014
New data							
Classroom observations	273 classrooms	273 classrooms				X	
Teacher survey	1,456 teachers	1,456 teachers		End of year		End of year	
Principal survey	91 principals	91 principals		End of year		End of year	
District interviews	14 district administrators			End of year		End of year	
Archival record data							
Student records (all students in study schools	14 districts			End of year		End of year	
Employee records for sampling (all teachers and principals in the study schools)	14 districts			January		January	
Employee records for tracking retention/mobility (all teachers and principals in the district)	14 districts				Novembe r		November
Performance evaluation system ratings (all treatment teachers and principals)	14 districts			End of year		End of year	

Classroom Observations

In addition to the classroom observations that will be conducted in treatment schools as a part of the evaluation system, we will conduct three independent observations within each of the 182 study schools, for a sample of 546 teachers representing the treatment and control groups in the spring of the second year of data collection (2013–14). These will include teachers in Grades 4–8.[11] Data from these observations will be used to conduct an analysis of the impact of the evaluation system on teachers' practices.

[11] These are the grade levels for which we expect to have access to current year and prior year test scores in reading and mathematics.

To enable the analyses of impact on teacher practice to complement the analyses of impact on student achievement, teachers in the independent sample of observations will be selected randomly, and stratified by grade-level and subject-area taught, from among the grade-levels and subjects for which student impacts will be analyzed (i.e., Grades 4–8, in mathematics and reading). The CLASS observation tool will be used for the independent observations in the districts using the CLASS observation tool as part of the study intervention, and the FFT observation tool will be used for the independent observations in the districts using the FFT observation tool as part of the study intervention.

Teacher and Principal Surveys

A stratified random sample of 16 teachers and one principal from each school across all study schools (treatment and control), for a sample of 2,912 teachers and 182 principals, will provide critical information on implementation and intermediate outcomes. Separate surveys will be designed and administered to teachers and principals. Both surveys, however, will include the following three domains. The first domain will include questions related to the implementation of the teacher and leader evaluation system and its components (to be given to only those individuals in the 91 treatment schools). The second domain will assess intermediate outcomes, such as participation in professional development, which may explain any impacts on student achievement (to be given to all respondents in both treatment and control schools, allowing for comparisons of responses between the two groups of schools). The third domain of questions will include demographic and background information (to be given to all respondents in both treatment and control schools; these data will be important to track because the composition of the teacher or principal force may change over time in response to the implementation of the initiative).

District Interviews

Semi-structured district interview protocols will be designed to serve multiple purposes. Interviewing one official from each participating district (e.g., a human resources administrator), one purpose will be to collect descriptive information related to district operations, such as the evaluation system currently used in the district and the features of the evaluation system as implemented by the district. Another purpose will be to document districts' experiences implementing the study's evaluation system, such as identifying which aspects of the system are challenging and how the district addresses those challenges, as well as documenting districts' perceptions related to the quality and usefulness of the evaluation system feedback results.

Archival Record Data Collection Protocols

The study team will develop protocols for obtaining archival record data from the study districts and from the technology platform used by the implementation team. Data collected will include the following: (1) student records linked to teachers, (2) employee records for sampling, (3) employee records for tracking retention/mobility, and (4) performance evaluation system ratings (for those in treatment schools).

Student Records for the Impact Analyses. To estimate program impact on student achievement, student records such as demographic and achievement information will be collected from all study districts. These data will be collected for each year of the implementation and for the year prior, as they become available in late spring or early summer in 2013 and 2014.

Employee Records for Sampling. To select our observation sample and survey sample, we will need to collect up-to-date school rosters prior to random sampling, in January of 2013 and January of 2014.

Employee Records for Tracking Mobility. To understand program impact on retention/mobility, we will gather archival data on teacher and leader school assignments and assignment fields (grade level and subject area) in November of 2013 and November of 2014. These data will be used to analyze district attrition (i.e., whether a teacher or principal remains in the district), school mobility (i.e., whether a teacher or principal has moved to another school in the district), and within-school assignment switching (i.e., whether a teacher has moved to a different assignment area within the same school—for example, from 4th grade to 2nd grade).

Performance Evaluation System Ratings. To understand performance ratings, we will gather archival data from the implementation team at the end of each academic year, in spring 2013 and spring 2014.

Supporting Statement for Paperwork Reduction Act Submission

A. Justification

1. Circumstances Making Collection of Information Necessary

Improving student achievement is a core priority for federal education policy. In 2009, only 33 percent of fourth graders and 32 percent of eighth graders performed at or above proficiency level in reading (National Center for Education Statistics, 2010b). Similarly, 39 percent of fourth graders and 34 percent of eighth graders attained the proficient level in mathematics (National Center for Education Statistics, 2010a). U.S. students' performance on international assessments lags far behind that of their international peers (Aud, Hussar, Kena, Bianco, Frohlich, et al., 2011; Hanushek, Peterson, & Woessmann, 2010).

In an effort to improve student outcomes, federal education policy—as manifested in the 2002 reauthorization of the Elementary and Secondary Education Act (ESEA), the American Reinvestment and Recovery Act (ARRA) of 2009, and the U.S. Department of Education's (ED's) Blueprint for the reauthorization of ESEA—targets teacher quality as a potential lever for improving student achievement. Several studies support this focus, with recent work examining individual teachers for several years to estimate the degree of variation in achievement gains that can be attributed to teachers rather than measurement error (see Bill & Melinda Gates Foundation, 2011; Schochet & Chiang, 2010). On balance, these studies suggest that being assigned to a teacher who is 1 standard deviation above average in effectiveness at raising student achievement may lead to a 0.1 standard deviation increase in student achievement. If such a teacher effect accumulates as students move from one grade to the next, it would translate into a substantial difference in achievement by the time students leave high school. Similar studies suggest that some variation in achievement gains can be attributed to principals, which indicates that principal quality may be another lever for policy makers (Hallinger & Heck, 1998; Leithwood et al., 2004; Leithwood & Jantzi, 2005; Leithwood et al., 2010; Waters, Marzano, & McNulty, 2003).

States and districts seeking to improve the measurement of teacher performance have pursued two main approaches. First, taking advantage of the greater availability of teacher-linked student achievement data, some have used student achievement growth as a teacher performance measure. At the same time, because many teachers teach grades or subjects that currently are not tested, and because of objections to relying solely on student growth measures (e.g., Koedel & Betts, 2009), some have pursued classroom observations for the measurement of teacher performance.

The federal government has invested considerable financial resources into supporting efforts to develop systems to measure educator effectiveness and use that information in human resource policies. The Teacher Incentive Fund (TIF) was established in 2006 with an initial funding base of approximately $100 million to identify and reward effective educators in high-poverty schools, based in part on student growth. The goal of the program is to improve student achievement by increasing the effectiveness of educators in high-poverty schools.

ARRA provided additional funding for TIF and established competitive grants to help states build their pool of effective teachers and principals through the Race to the Top (RTT) program, in which a core priority area is the effectiveness of teachers and leaders. Using RTT grants, states and districts are developing new teacher and leader evaluation systems, with the goal of improving teacher effectiveness and student outcomes. The U.S. Department of Education's (ED's) Blueprint for the reauthorization of ESEA proposes a continued emphasis on the development of such systems. ARRA included $4.35 billion in funds for RTT.

Although states and districts across the country have been developing and implementing new evaluation systems to identify and reward highly effective educators, no large-scale study has yet been carried out to confirm the efficacy of these new systems. The Teacher and Leader Evaluation Systems (TLES) Study will address this need, through an experimental design that provides a comprehensive teacher and leader evaluation system (including teacher observation, principal leadership evaluation, and student growth components) to assess impacts on instructional practice, principal leadership, and student achievement. For these reasons, we request initial clearance for the study's district-level screening protocol used for recruitment.

2. Purposes and Uses of the Data

This request focuses on the collection of data from school districts to determine their suitability for recruitment into the study. To enable a meaningful experimental test of the study's teacher and leader evaluation system, the districts recruited into the study will have to be engaged in traditional practices and meet other criteria that make them suitable for recruitment.

Specifically, the district screening interviews will ascertain the nature of eligible districts' current teacher and leader evaluator system. We will ask districts how teachers and leaders are evaluated, what observation tools are used, and what commercial products if any are being used. We also need to know who currently conducts those observations, how they are trained, and how often observations occur for all teachers. These are system features that distinguish districts using traditional systems from districts using more advanced systems.

The district screening interviews also will ascertain districts' data capacity for measuring student growth. We will investigate the extent to which the district can link student course-taking and test data with teacher information, and what type of value-added or other student growth analyses are being conducted by the district.

Finally, information from the screening interviews will be used for further discussions with the district, including speaking with data specialists during visits to the district.

3. Use of Technology to Reduce Burden

District screening interviews will be conducted by phone. For purposes of gathering the information needed to determine district eligibility for the study, telephone interviews have many advantages over mail surveys. First, a telephone interview is less burdensome for respondents, who can provide oral answers. Consequently, a telephone interview is likely to yield a better response rate than a paper survey. Second, telephone interviews can generate responses within minutes once the interviewer reaches the respondent, which helps to maximize the efficiency of our district screening and recruitment process. Third, the interviewer can immediately probe for further information to clarify ambiguous or conditional responses.

4. Efforts to Identify Duplication

Before administering the screening protocol, every effort will be made to collect the needed information from archival data on state policy, the Common Core of Data (CCD), district websites, and other knowledgeable sources. These efforts are described in detail under B.1 (Respondent Universe and Sampling Methods) in the subsection called, Identifying the Pool of Districts to Be Screened.

Although these sources will help AIR target its efforts, much of the information required to identify eligible districts and schools is either not publicly available or is not kept up-to-date. The screening interviews will therefore allow study staff to collect information not available elsewhere and verify information gathered from public sources.

5. Methods to Minimize Burden on Small Entities

To be considered a small entity by OMB, a school district would need to have a population of fewer than 50,000 students. Our criteria will exclude some of the smallest districts in selecting the district sample. Specifically, as explained further under B.1 (Respondent Universe and Sampling Methods) in the subsection called, Identifying the Pool of Districts to Be Screened, our criteria will require that districts (1) have at least 10 qualifying schools, (2) have data systems that support value-added modeling, and (3) have not announced a recent implementation (or planned implementation) of a new teacher or leader evaluation system, to take effect before 2014-15. These criteria will exclude some of the smallest districts that might be the most burdened by the study requirements.

6. Consequences of Not Collecting the Data

The TLES Study represents an ongoing effort by the Department of Education to rigorously study the effects of teacher and leader evaluation systems on educators and students. States and districts are increasingly reshaping teacher and leader evaluation systems, and without this study, states and districts will have a limited understanding of how these systems affect their students and educators.

With regard to this initial clearance package, measuring the impact of our proposed evaluation system depends on our ability to find districts with current evaluation systems that serve as a stark contrast to our own system. If we do not gather this information, then our control schools may have evaluation systems very similar to the system provided to the treatment schools and thus render the treatment-control comparisons less meaningful.

7. Special Circumstances

No special circumstances apply to this study.

8. Federal Register Comments and Persons Consulted Outside the Agency

A 60-day notice was published in the Federal Register, volume 76, page 73,606 on Tuesday, November 29, 2011, providing an opportunity for public comments. No public comments were received. A 30-day notice will be published to further solicit comments.

To assist with the development of the screening criteria and the study as a whole, project staff will draw on the experience and expertise of a network of outside experts who will serve as our technical working group members.

Exhibit 5. Expert Reviewers

Proposed Expert Reviewer	Professional Affiliation	Area(s) of Expertise
Thomas Cook	Northwestern University	Impact evaluations
Laura Goe	Educational Testing Services	Teacher performance evaluation models
Thomas Dee	University of Virginia	Performance compensation
Daniel McCaffrey	Rand Corporation	Value-added measurement
Catherine McClellan	Clowder Consulting	Psychometrics; measures of instructional practice
Jonah Rockoff	Columbia University	Teacher performance and student achievement
Patrick Schuermann	Vanderbilt University	Teacher and principal performance evaluation
Carla Stevens	Houston Independent School District	Educator Quality
John Tyler	Brown University	Teacher performance and student achievement
Judy Wurtzel	Independent Consultant	Educator quality

9. Payment or Gifts

The administration of the district-level screening protocol and completion of recruitment activities will involve no payments or gifts. In the addendum package, we will specify incentives that are intended to encourage cooperation with the study's data collections.

10. Assurances of Confidentiality

No confidential data will be sought during the recruitment phase of the study, for which clearance is being sought in this package.

The following statement applies to procedures to take place during the data collection phase of the study, for which clearance will be sought in a separate OMB submission. A consistent and cautious approach will be taken to protect all information collected during the data collection phase of the study. This approach will be in accordance with all relevant regulations and requirements. These include the Education Sciences Institute Reform Act of 2002, Title I, Part E, Section 183, which requires "[a]ll collection, maintenance, use, and wide dissemination of data by the Institute … to conform with the requirements of section 552 of Title 5, United States Code, the confidentiality standards of subsections (c) of this section, and sections 444 and 445 of the General Education Provisions Act (20 U.S.C. 1232 g, 1232h).” These citations refer to the Privacy Act, the Family Education Rights and Privacy Act, and the Protection of Pupil Rights Amendment. In addition, for student information, the project director will ensure that all individually identifiable information about students, their academic achievements and families, and information with respect to individual schools shall remain confidential in accordance with section 552a of Title 5, United States Code, the confidentiality standards subsection (c), and sections 444 and 445 of the General Educations Provision Act.

Subsection (c) of Section 183, referenced above, requires the director of IES to “develop and enforce standards designed to protect the confidentiality of persons in the collection, reporting, and publication of data.” The study will also adhere to requirements of subsection (d) of Section 183 prohibiting disclosure of individually identifiable information as well as making the publishing or inappropriate communication of individually identifiable information by employees or staff a felony.

AIR will use the information collected in the study for research purposes only. When reporting the results, data will be presented only in aggregate form, such that individuals and institutions will not be identified. A statement to this effect will be included with all requests for data. All members of the study team with access to the data will be trained and certified on the importance of privacy and data security. All data will be kept in secured locations and identifiers will be destroyed as soon as they are no longer required.

The following safeguards are routinely employed by AIR to carry out privacy assurances during the study:

- All AIR employees sign a privacy pledge emphasizing its importance and describing their obligation.
- Identifying information is maintained on separate forms and files, which are linked only by sample identification number.
- Access to hard copy documents is strictly limited. Documents are stored in locked files and cabinets. Discarded materials are shredded.
- Computer data files are protected with passwords and access is limited to specific users.
- Especially sensitive data are maintained on removable storage devices that are kept physically secure when not in use.

11. Justification of Sensitive Questions

No questions of a sensitive nature will be included in the screening protocol.

12. Estimates of Hour Burden

There are two components for which we have calculated hours of burden for this clearance package: the district-level screening and the follow-up recruitment activities.

The total estimated hour burden to screen districts for eligibility for the TLES study is 51 hours, which includes time for 85 percent of the 120 district officials in the 120 candidate districts to respond to a 30-minute district-level screening protocol.

The total estimated hour burden to complete follow-up recruitment activities and recruitment site visits is 840 hours. The burden estimate for recruitment will vary greatly with the district's persistence in the pool of potential candidates. The averaged burden estimate for recruitment of 24 hours per district includes time for all 45 viable districts to read recruitment materials and participate in a follow-up telephone call and time to progressively pursue smaller numbers of districts to participate in further follow-up calls, to host site visits, to attend a (voluntary) recruitment conference, and to negotiate final agreements.

On the basis of average hourly wages of participants, the cost of the recruitment efforts amount to $40,095 for screening and recruitment combined. Exhibit 6 summarizes the estimates of respondent burden for these two study activities. Burden estimates for other data collection activities will be included in the second OMB package.

Exhibit 6. Hour Burden for Respondents

Task	Total Sample Size	Estimated Response Rate	Number of Responses	Time Estimate Per District (in hours)	Number of Admini-strations	Total Hours	Hourly Rate	Estimated Monetary Cost of Burden
District Screening	120	85%	102	0.5	1	51	$45	$2,295
Follow-up Recruitment Activities	45	100%	45	8	1	360	$45	$16,200
Recruitment Site Visits	30	100%	30	16	1	480	$45	$21,600
Total			177			891		$40,095
Total Annual			59			297		$13,365

13. Estimate of Cost Burden to Respondents

There are no additional respondent costs associated with this data collection other than the hour burden accounted for in item 12.

14. Estimate of Annual Cost to the Federal Government

The estimated cost for all aspects of the study is $16,783,022 over five years, making the annual cost to the federal government $3,356,604. Total Year 1 cost is $4,668,189; Year 2 cost is $5,998,265; Year 3 cost is $5,224,346; Year 4 cost is $765,151; and the Year 5 cost is $127,071.

15. Program Changes or Adjustments

This request is for a new information collection.

16. Plans for Tabulation and Publication of Results

There will be no formal tabulations or reports based on the district screening. This information will be used for internal purposes only.

Findings from the TLES Study will be reported to IES by AIR in two substantive reports. The first report will document the results of the data analyses conducted on the data collected for the

2012–13 school year, including mobility as measured in fall 2013. This report will include a description of the study design and results from both descriptive analyses and impact analyses. More specifically, the descriptive analyses will include the following:

- Descriptive analyses of the characteristics of the study sample
- Analyses of the equivalence of the treatment and control groups in their background characteristics after randomization
- Descriptive analysis of the fidelity of implementation of the interventions that are part of the teacher and leader evaluation system

In addition, the report will provide results on the effects of the interventions on both teacher and principal mobility and student achievement during the first treatment year (2012–13). These analyses will be carried out using hierarchical linear modeling to take into account the nesting of students and teachers within schools, and will incorporate covariates measured at baseline to maximize precision. To avoid potential selection bias, the impact analyses will employ an "intent-to-treat" approach, in which all teachers and principals in all randomly assigned schools during the 2012–13 school year are included in the analyses, whether or not the teachers or principals actually participated in the intended teacher and leader evaluation system or participated to the full extent expected.

The final report will be a capstone report summarizing the entire project and its results. The main focus of the report will be the results pertaining to the cumulative effects of providing the intervention for two years (i.e., 2012–13 and 2013–14) on teacher and principal mobility (RQ4), and both instructional practice (RQ5) and student achievement (RQ6). As in the first report, these analyses will be conducted using hierarchical linear modeling with covariate adjustment. In addition, the report will examine the possible relationships between teacher/principal experience and the impact of the interventions; the relationship between student characteristics and impact; and the cost of the intervention. The report also will include a comprehensive analysis of teacher, principal, and student mobility and its potential effects on study results. The timeline for the dissemination of the study results is summarized in Exhibit 7.

Exhibit 7. Schedule for Dissemination of Study Results

Deliverable	Anticipated Release Date
First Report	December 2014
Final Report	December 2015

17. Approval to Not Display OMB Expiration Date

Approval is not being requested; all data collection instruments will include the OMB expiration date.

18. Explanation of Exceptions

No exceptions are requested.

References

Aud, S., Hussar, W., Kena, G., Bianco, K., Frohlich, L., Kemp, J., et al. (2011). *The condition of education 2011* (NCES 2011-033). Washington, DC: U.S. Department of Education, National Center for Education Statistics. Retrieved November 4, 2011, from http://nces.ed.gov/programs/coe/pdf/coe_msl.pdf

Bill & Melinda Gates Foundation. (2011). *Learning about teaching: Initial findings from the measures of effective teaching project.* Seattle, WA: Author. Retrieved November 4, 2011, from http://www.metproject.org/downloads/Preliminary_Findings-Research_Paper.pdf

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis, 29*(1), 30–59.

Condon, C., & Clifford, M. (2010). *Measuring principal performance: How rigorous are commonly used principal performance assessment instruments?* Naperville: Learning Point Associates.

Danielson, C. (2010). Evaluations that help teachers learn. *Educational Leadership, 68*(4), 35–39.

Danielson, C., & McGreal, T. L. (2000). *Teacher evaluation: To enhance professional practice.* Alexandria, VA: Association for Supervision and Curriculum Development.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal, 38,* 915–945.

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness.* Washington, DC: National Comprehensive Center for Teacher Quality.

Goe, L., Biggers, K., & Croft, A. (2011). Linking teacher evaluation to professional development: Focusing on improving teaching and learning. Presentation at the National Comprehensive Center for Teacher Quality conference *Enhancing Teacher Evaluation: A Critical Lever for Improving Teaching and Learning,* Washington, DC, May 10–11.

Goldring, E., Carvens, X., Murphy, J., Porter, A., Elliott, S., & Carson, B. (2009). The evaluation of principals: What and how do states and urban districts assess leadership? Elementary School Journal, 110(1), 19–39.

Hallinger, P., & Heck, R. H. (1998). Exploring the principal's contribution to school effectiveness: 1980–1995. *School Effectiveness and School Improvement, 9*(2), 157–191.

Hanushek, E. A., Peterson, P. E., & Woessmann, L. (2010). *U.S. math performance in global perspective: How well does each state do at producing high-achieving students?* (PEPG Report No. 10-19). Cambridge, MA: Harvard Kennedy School, Taubman Center for State and Local Government, Program on Education Policy and Governance & Education Next. Retrieved November 4, 2011, from http://www.hks.harvard.edu/pepg/PDF/Papers/PEPG10-19_HanushekPetersonWoessmann.pdf.

Jacob, R., Zhu, P., & Bloom, H. S. (2009). *New empirical evidence for the design of group randomized trials in education.* New York: MDRC.

Leithwood, K., Harris, A., Strauss, T. (2010). *Leading school turnarounds: How successful leaders transform low-performing public schools.* San Francisco: Jossey-Bass.

Leithwood, K., & Jantzi, D. (2005). A review of transformational school leadership research 1996–2005. *Leadership and Policy in Schools, 4*(3), 177–199.

Leithwood, K., Louis, K. S., Anderson, S., & Wahlstrom, K. (2004). *How leadership influences student learning*. St. Paul: University of Minnesota, Center for Applied Research and Educational Improvement.

Murray, D. M., Varnell, S. P., Blitstein, J. L. (2004). Design and analysis of group randomized trials: a review of recent methodological developments. *American Journal of Public Health, 94,* 423–432.

National Center for Education Statistics. (2010a). *The nation's report card: Mathematics 2009* (NCES 2010–451). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. Retrieved November 4, 2011, from http://nces.ed.gov/nationsreportcard/pdf/main2009/2010451.pdf

National Center for Education Statistics. (2010b). *The nation's report card: Reading 2009* (NCES 2010–458). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. Retrieved November 4, 2011, from http://nces.ed.gov/nationsreportcard/pdf/main2009/2010458.pdf

Porter, A. C., Youngs, P., & Odden, A. (2001). *Advances in teacher assessment and their uses. In V. Richardson (Ed.), Handbook of research on teaching,* 4th ed. (259–297). Washington, DC: American Educational Research Association.

Schochet, P. (2005). *Statistical power for random assignment evaluations of education programs.* Princeton, NJ: Mathematica Policy Research.

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

Schochet, P. Z., & Chiang, H. S. (2010). *Error rates in measuring teacher and school performance based on student test score gains* (NCEE 2010-4004). Washington, DC:

U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Retrieved November 4, 2011, from http://eric.ed.gov/PDFS/ED511026.pdf

Siddiqui, O., Hedeker, D., Flay, B. R., & Hu, F. B. (1996). Intraclass correlation estimates in a school-based smoking prevention study. *American Journal of Epidemiology 144*(4), 425–433.

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.

U.S. Department of Education. (2010). *A blueprint for reform: The reauthorization of the Elementary and Secondary Education Act*. Washington, DC: Author. Retrieved July 8, 2011, from http://www2.ed.gov/policy/elsec/leg/blueprint/blueprint.pdf

Waters, T., Marzano, R. J., & McNulty, B. (2003). *Balanced leadership: What 30 years of research tells us about the effect of leadership on student achievement.* Aurora, CO: Mid-continent Research for Education and Learning. Retrieved November 4, 2011, from http://www.mcrel.org/pdf/LeadershipOrganizationDevelopment/5031RR_BalancedLeadership.pdf

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on teacher effectiveness.* New York: The New Teacher Project.

Weiss, E. M., & Weiss, S. G. (1998). *New directions in teacher evaluation* (ERIC Digest). Washington, DC: ERIC Clearinghouse on Teaching and Teacher Education.