DOES TEACHER EVALUATION IMPROVE SCHOOL PERFORMANCE? EXPERIMENTAL EVIDENCE FROM CHICAGO'S EXCELLENCE IN TEACHING PROJECT

Matthew P. Steinberg
(corresponding author)
Graduate School of Education
University of Pennsylvania
Philadelphia, PA 19104
steima@gse.upenn.edu

Lauren Sartain
Consortium on Chicago School Research
Harris School of Public Policy
University of Chicago
Chicago, IL 60622
lsartain@uchicago.edu

Abstract

Chicago Public Schools initiated the Excellence in Teaching Project, a teacher evaluation program designed to increase student learning by improving classroom instruction through structured principal–teacher dialogue. The pilot began in forty-four elementary schools in 2008–09 (cohort 1) and scaled up to include an additional forty-eight elementary schools in 2009–10 (cohort 2). Leveraging the experimental design of the rollout, cohort 1 schools performed better in reading and math than cohort 2 schools at the end of the first year, though the math effects are not statistically significant. We find the initial improvement for cohort 1 schools remains even after cohort 2 schools adopted the program. Moreover, the pilot differentially impacted schools with different characteristics. Higher-achieving and lower-poverty schools were the primary beneficiaries, suggesting the intervention was most successful in more advantaged schools. These findings are relevant for policy makers and school leaders who are implementing evaluation systems that incorporate classroom observations.

doi:10.1162/EDFP_a_00173

1. INTRODUCTION

One of the most persistent and urgent problems facing education policy makers is the provision of highly effective teachers in all of our nation's classrooms. The increasing demand for high-quality teachers, particularly in urban public schools and in areas such as mathematics and science education, has been well documented for at least three decades (NCEE 1983; Ingersoll 2001; Murnane and Steele 2007). Indeed, of all school-level factors related to student learning and achievement, the student's teacher has consistently been shown to be the most important (Goldhaber 2002; Rockoff 2004; Rivkin, Hanushek, and Kain 2005). Even with substantial within-school variation in teacher effectiveness (Rivkin, Hanushek, and Kain 2005; Aaronson, Barrow, and Sander 2007), historically, teacher evaluation systems have inadequately differentiated teachers who effectively improve student learning from lower-performing teachers. Indeed, a recent study by The New Teacher Project (TNTP) found that more than 99 percent of teachers were rated satisfactory in districts that use binary evaluation ratings ("satisfactory" or "unsatisfactory" ratings as the mutually exclusive choices available to school principals and administrators; Weisberg et al. 2009).

Only recently have policy efforts begun to address alternative methods for evaluating teacher performance. Increasingly, state and local education agencies are replacing traditional teacher evaluation approaches in order to incorporate multiple methods of assessing and evaluating teachers. According to the National Council on Teacher Quality, between 2009–10 and 2011–12, thirty-two states and the District of Columbia altered their teacher evaluation policies. This move has been influenced, in large part, by the federal Race to the Top (RTTT) competition. The 2010 RTTT competition emphasized more rigorous performance evaluation through the use of multiple measures of teacher performance as well as the incorporation of multiple teacher ratings categories to differentiate teacher effectiveness. A prominent feature of performance evaluation systems supported by RTTT is the use of student test score data to estimate a teacher's idiosyncratic contribution to student learning (so called "value-added" metrics). Nevertheless, even as teacher evaluation systems increasingly incorporate student test score data, a majority of teachers—upward of 70 percent nationally—teach in grades or subjects in which standardized achievement exams are not administered, and therefore will not have a value-added score (Watson, Kraemer, and Thorn 2009). As a result, qualitative, classroom-observation-driven measures of teacher performance remain critically important components of teacher ratings. For example, in Tennessee, one of the first states awarded an RTTT grant, half of a teacher's overall evaluation score under the Tennessee

Educator Acceleration Model is based on classroom observation and teacher conferences.[1]

Value-added and student-growth measures based on test scores and classroom-based observation data capture very different dimensions of a teacher's performance. Value-added metrics and other measures that rely on student achievement data are output-based and represent ex post measures of a teacher's performance in the classroom. It is not clear, a priori, that a teacher evaluation system based primarily on student achievement should lead to improvements in teacher performance and, ultimately, student achievement because such metrics do little (if anything) to alter a teacher's instructional practice. Indeed, we are unaware of any evidence in the research literature that provides a causal link between an evaluation system based primarily on student test scores and student academic achievement.

On the other hand, classroom observations of a teacher's instructional practice coupled with principal–teacher conferences represent process-oriented measures (Goe, Bell, and Little 2008). These measures are designed to capture the quality of within-classroom interactions among students and teachers in the context of daily instruction. The theory of action embedded in such process-based systems is that changes to teacher practice through an iterative process of observation and conferencing—all focused on improving lesson planning and preparation, the classroom environment, and instruction—should lead to direct changes in student performance. As teachers refine their practice and target the learning needs of their students, student performance should improve.

The emphasis on process-based measures of teacher practice in newly developed evaluation systems recognizes an important fact about teacher human capital: Namely, that the static dimension of teacher human capital—immutable teacher characteristics such as demographics (race, gender, age), prior experience (years in the teaching profession and tenure status), educational attainment (master's degree), and observed measures of cognitive achievement (college selectivity, test scores on the SAT/ACT and state certification exams, college grades)—plays a limited role in improving student achievement. Indeed, evidence suggests that the highest degree a teacher earns (usually a master's degree) makes little to no difference for student achievement (Clotfelter, Ladd, and Vigdor 2007; Goldhaber 2007), and there is little to no relationship between student performance and college selectivity, test scores, or college grades (Goldhaber 2002, 2007; Harris and Sass 2007; Kane, Rockoff, and Staiger 2008). Although there is consistent support

1. Source: Tennessee Department of Education (see http://team-tn.org/).

for positive returns to teacher experience, those returns are concentrated in a teacher's first two or three years in the classroom (Goldhaber 2002; Rivkin, Hanushek, and Kain 2005; Clotfelter, Ladd, and Vigdor, 2007). Given the limited evidence on the relationship between the static dimension of teacher human capital and student achievement, it is no surprise that policy makers' and school leaders' ability to identify high-quality teachers at the time of hire is limited (Rockoff et al. 2011), particularly because these observable dimensions of teachers only account for approximately 3 percent of the total variation in student achievement (Goldhaber 2002).

If policy makers are unable to substantively move student outcomes by targeting the static dimension of teacher human capital, what about interventions that aim to influence the actual practice of teaching? Recent evidence from Cincinnati Public Schools suggests that process-oriented measures of teacher effectiveness, captured by the Danielson Framework for Teaching (Danielson 1996), promote student achievement growth in math both during the school year in which the teacher is evaluated as well as in the years after evaluation (Kane et al. 2011; Taylor and Tyler 2012).[2]

In this paper we explore the Excellence in Teaching Project (the "pilot," or EITP), a teacher evaluation system based on the Danielson Framework implemented in Chicago Public Schools (CPS) during the 2008–09 school year. Among a sample of elementary schools, forty-four schools were randomly assigned to implement EITP in 2008–09 (cohort 1), and an additional forty-eight elementary schools implemented EITP for the first time during the 2009–10 school year (cohort 2). We leverage both the experimental design of the pilot in year 1 as well as the timing of program implementation among cohort 2 schools in year 2 to answer the following questions:

(1) What effect did the pilot teacher evaluation system have on school-level performance in mathematics and reading?
(2) Did the pilot teacher evaluation system differentially impact schools with different characteristics (for example, did the pilot program have a greater impact on lower or higher achieving schools)?
(3) Did the effect of the pilot, if any, persist over time?

We find that at the end of the first year (2008–09) of the pilot, cohort 1 schools performed better in reading and math than cohort 2 schools, although only the reading effects are statistically significant. We also find persistent gaps in reading and math achievement between cohort 1 and cohort 2 schools

2. Taylor and Tyler (2012) do not find statistically significant differences in student reading achievement associated with a teacher's participation in the teacher evaluation system.

after the second and third years of the pilot (the 2009–10 and 2010–11 school years, respectively). Moreover, at the end of the first year, the teacher evaluation pilot differentially impacted schools with different observable characteristics. Specifically, pilot schools with higher pretreatment student achievement performed better in reading than pilot schools with lower pretreatment student achievement, controlling for initial achievement levels. Further, higher-poverty pilot schools performed worse than lower-poverty schools in both math and reading, controlling for initial poverty levels.

We begin by describing the EITP pilot. Then we discuss the nature and implementation of teacher evaluation in Chicago, attending to the policy context that likely shaped our empirical findings. We then describe the data and the empirical methods used to estimate the school achievement effects. Finally, we present and discuss the findings and conclude.

2. TEACHER EVALUATION IN CHICAGO PUBLIC SCHOOLS

For nearly four decades prior to the introduction, in 2008, of the Excellence in Teaching Project, CPS teachers were observed and evaluated based on a checklist of classroom practices.[3] The checklist was organized into three sections (and included 19 practices): (1) Instruction, (2) School Environment, and (3) Professional and Personal Standards. During a classroom observation of a teacher's lesson, the observer (either the principal or assistant principal, though the principal was primarily the observer in the CPS context) would check one of three boxes (Strength, Weakness, Does not apply) next to each of the practices. The checklist was unpopular among both teachers and principals. High-performing teachers believed the system did not provide meaningful feedback on their instruction, and only 39 percent of veteran principals agreed that the checklist allowed them to adequately address teacher underperformance (TNTP 2007; Sartain et al. 2011). In particular, no formal guidance (in the form of a rubric) was provided to either party on what constituted, for example, the strong or weak *application of contemporary principles of learning theory and teaching methodology* (one of the instructional practices listed on the checklist). Moreover, there was no formal correspondence between a teacher's ratings on the checklist and his/her final summative evaluation rating, which determined teacher tenure. One major consequence of this checklist-based evaluation process was little differentiation among teachers in terms of their summative performance evaluations. Nearly all teachers (93 percent) received performance evaluation ratings of "Superior" or "Excellent" (based on a

3. Under the checklist system of teacher evaluation, tenured teachers rated excellent or superior were rated every two years (rather than annually), and probationary (nontenured) teachers were evaluated annually (TNTP 2007). See Sartain et al. (2011) for a copy of the checklist as well as the complete Danielson Framework.

four-tiered rating system) while at the same time 66 percent of CPS schools failed to meet state proficiency standards under Illinois's accountability system (TNTP 2007).

3. THE EXCELLENCE IN TEACHING PROJECT

As a result of dissatisfaction with what many perceived to be an ineffective evaluation system, beginning in 2006 the EITP was developed through a joint partnership with leadership from CPS under then-CEO Arne Duncan and the Chicago Teachers Union. The joint committee of CPS and the Chicago Teachers Union met together over the next two years to negotiate the details of the teacher evaluation pilot. In the summer of 2008, just prior to implementation, the district and union disagreed about how the pilot evaluation system would impact the evaluation ratings of nontenured teachers. As a result, while the district moved forward with the pilot, the classroom observation ratings teachers received under the EITP could not be used for teacher accountability, such as tenure decisions. Even without union support, the district wanted to use the classroom observation process as a means of formative, ongoing assessment for teachers, providing them with structured feedback on their instructional practices. The district's stated goals of the pilot were to improve teaching and learning, develop a stronger professional learning climate, and foster a constructive climate around teacher evaluation. As a result, EITP was specifically focused on the aspect of personnel evaluation related to professional development and teacher growth. Given these goals, any observed impacts of the new teacher evaluation pilot on school performance will likely operate through these mechanisms: (1) increased principal capacity as instructional leader through district efforts to support and train principals around this initiative; (2) improvements in the instructional quality within the classroom through principal feedback to teachers during the classroom observation and conferencing process; and (3) a more coherent school learning climate, particularly the extent of collaboration between principals and teachers through the nature of the conferencing process.[4]

CPS selected four (of the seventeen) elementary school instructional areas in which to implement the pilot evaluation system.[5] The areas that the

4. In ongoing work, the authors are exploring whether the teacher evaluation pilot led to improvements in these intermediate outcomes.
5. The elementary school areas were primarily geographic-based subdistricts of CPS. At that time, an elementary school instructional area worked with designated schools on their instructional practices and student services with the goal of improving student learning. Area instructional offices played a major role in the planning and implementation of various programs and initiatives designed to improve and enhance academic achievement. Since the initial implementation of the pilot, however, CPS has twice overhauled these offices, changing the schools that constitute the areas and altering the offices' role in supporting schools. The area offices now tend to focus on performance management and data usage.

district targeted for this pilot were in different parts of the city and selected to represent the district as a whole. The pilot consisted of two main features: principal training and the classroom observation and conference process itself. A key to the pilot was its focus on building principal-level capacity to implement this new evaluation process, which represented a dramatic departure from the checklist. We first consider the process by which principals observed and conferenced with teachers. Then, we discuss the year-by-year training and district-level support principals received as part of the school-level implementation of EITP.

Classroom Observation and Principal–Teacher Conferences

The classroom observation process occurred formally two times per year for all teachers, irrespective of tenure status, as part of the district–union teacher contract.[6] As part of the classroom observation and conference process, principals and teachers first engaged in a brief (15–20 minute) pre-observation conference during which they reviewed the observation rubric. The conference also gave the teacher an opportunity to share any information about their classroom with the principal, such as specific issues with individual students or areas of practice about which the teacher wanted detailed feedback. Then, the classroom observation occurred, and the observation period was supposed to cover a 30–60 minute instructional unit or lesson. During this time, the principal was to take detailed notes about what the teacher and students were doing. The training for principals emphasized the collection of evidence, rather than opinions, about what was happening in the classroom. After the observation, the principal was expected to match his or her classroom observation notes to the Danielson Framework rubric in order to rate teacher performance in ten areas of instructional practice.

The Danielson Framework delineates four levels of performance (unsatisfactory, basic, proficient, and distinguished) across four domains: (1) Planning and Preparation; (2) The Classroom Environment; (3) Instruction; and (4) Professional Responsibilities. The EITP focused on just two of these domains—The Classroom Environment and Instruction—each of which included five areas of classroom practice. Table 1 lists the ten areas of instruction

6. For both cohort 1 and cohort 2 schools, 100 percent of teachers, both tenured and nontenured, were subject to two yearly classroom observations during the first year of EITP implementation. Regardless of tenure status, principals were required to observe and conference with teachers two times per year. However, nontenured teachers received an official summative evaluation rating every year, and tenured teachers received an official summative evaluation rating every other year. The 2008–09 school year was an off-year for official tenured teacher evaluation, though the teacher contract still required the principal to conduct classroom observations and provide feedback to tenured teachers. In 2009–10, all teachers across the district received an official summative evaluation rating.

Table 1. Components of Charlotte Danielson's Framework for Teaching

Domain 2: The Classroom Environment	Domain 3: Instruction
Creating an Environment of Respect and Rapport	Communicating with Students
Establishing a Culture for Learning	Using Questioning and Discussion Techniques
Managing Classroom Procedures	Engaging Students in Learning
Managing Student Behavior	Using Assessment in Instruction
Organizing Physical Space	Demonstrating Flexibility and Responsiveness

Table 2. Example of a Rubric for One Danielson Component: Using Questioning and Discussion Techniques

Level of Performance/ Rating	General Description	Specific Rubric for 3B: Using Questioning and Discussion Techniques
Unsatisfactory	Teaching is below the standard of "do no harm" and requires immediate intervention.	Teacher's questions are low-level or inappropriate, eliciting limited student participation and recitation rather than discussion.
Basic	Teacher understands the components of teaching, but implementation is sporadic.	Some of the teacher's questions elicit a thoughtful response, but most are low-level, posed in rapid succession. Teacher attempts to engage all students in the discussion are only partially successful.
Proficient	Teacher has mastered the work of teaching.	Most of the teacher's questions elicit a thoughtful response, and the teacher allows sufficient time for students to answer. All students participate in the discussion, with the teacher stepping aside when appropriate.
Distinguished	Teacher has established a community of learners with students assuming responsibility for their own learning.	Questions reflect high expectations and are culturally and developmentally appropriate. Students formulate many of the high-level questions and ensure that all voices are heard.

in which principals rated teachers in Chicago, and table 2 provides an example of the rubric for one of the rating areas under the Instruction domain, *Using Questioning and Discussion Techniques*. Within a week of the observation, the principal and teacher conducted a post-observation conference. During the conference, the principal shared evidence from the classroom observation, as well as the Danielson ratings, with the teacher. Principals and teachers were expected to discuss any areas of disagreement in the ratings, with a specific focus on ways to improve the teacher's instructional practice and, ultimately,

student achievement. Evidence suggests that the principal ratings were both reliable and valid measures of teacher practice (Sartain et al. 2011).

Principal Training and Support

EITP represented a dramatic shift in the way teacher evaluation had occurred in CPS, and central office staff sought to develop principals' capacity to conduct these classroom observations and conferences. Table 3 summarizes, by year, the extent of principal training and district support and oversight for EITP implementation. In 2008–09, the first year of implementation, forty-four participating principals received approximately 50 hours of training and support, with three days of initial training during the summer before implementation and follow-up sessions throughout the school year (Sartain, Stoelinga, and Krone 2010). The content of the training for principals included the use of the Danielson Framework to rate teaching practice, methods for collecting evidence, and best practices for conducting classroom observations. The training also included support for principals in coaching teachers, though the primary focus was on the rating process. The follow-up sessions consisted of seven monthly meetings in which principals discussed a variety of implementation issues in the context of professional learning communities consisting of other participating principals. During the professional learning community time, principals brought materials from classroom observations which they had conducted, and engaged in small group discussion with their colleagues, providing a rich set of supports for principals as they implemented EITP for the first time.

Principals also received four half-day trainings during the school year, which provided an opportunity for them to update their understanding and use of the rubric for evaluating teachers. The implementation of EITP relied heavily on principals and their capacity to work with their teachers. Principals informed teachers about the new observation process, and observed and conferenced with teachers to provide targeted guidance on teachers' instructional practice. In doing so, principals required committed support from the CPS central office. During this first year of implementation, central office administrators responsible for EITP engaged with principals through weekly e-mails, providing consistent reminders to principals about observation deadlines and other EITP requirements. Moreover, principals could request time with EITP central office staff to review their teacher ratings as a means of calibrating their observation sessions to EITP central office expectations. Finally, principals received individualized ratings reports from the University of Chicago Consortium on Chicago School Research (CCSR). The CCSR reports provided principals with a comparison of their own teacher ratings to ratings generated by trained external observers of the same teachers. These reports supported

Table 3. Summary of EITP Principal Training and Support

2008–09		2009–10		2010–11	
Principal Training					
Cohort 1	**Cohort 2**	**Cohort 1**	**Cohort 2**	**Cohort 1**	**Cohort 2**
• Introductory all-principal meeting with central office leadership • Three-day summer training prior to implementation • Four half-day refresher trainings throughout the school year • Seven monthly professional learning community meetings with participating principals	• None—not yet implementing EITP	• Two professional development refresher days; one in the fall and a second in the spring	• Two days of initial training; one in spring and a second in the summer prior to implementation	• None—program was voluntary	
District Support and Oversight of Principals					
• Weekly e-mails to principals reminding them of observation deadlines and other EITP requirements • Two individualized ratings reports for principals from CCSR • Principals could request individualized ratings calibration sessions with central office staff	• None—not yet implementing EITP	• Principals could request technical assistance from central office as needed	• Principals could request technical assistance from central office as needed	• None—program was voluntary	

Notes: There were 44 cohort 1 schools first implementing EITP in the 2008–09 school year and 48 cohort 2 schools first implementing EITP in the 2009–10 school year.
Source: Sartain et al. (2011) and Sheila Cashman, personal communication (17 June 2013).

principals in making adjustments to their own ratings of teacher performance (Sartain et al. 2011).

In 2009–10, the forty-four cohort 1 schools continued to participate in EITP, and an additional forty-eight schools (cohort 2) implemented EITP for the first time. Whereas cohort 1 school principals received extensive training, as described here, the extent of principal training and support for the forty-eight new schools differed dramatically from the early adopters. In their first year of adopting EITP, cohort 2 principals received two days of initial training—one in spring 2009 and a second in summer 2009—on how to collect evidence of best teaching practices during classroom observations and how to rate these practices using the Danielson Framework. Cohort 2 principals also received significantly less district-level support throughout the school year than cohort 1 principals in their first year of implementation. Although cohort 2 principals could request technical assistance from EITP central office staff, these principals did not receive the type of ongoing technical support and oversight that cohort 1 principals received during their first year of implementation. Indeed, cohort 1 principals received the same level of support and ongoing training in their second year of implementation as did the cohort 2 principals in their first year. In conversations with EITP central office staff, one staff member indicated that between the 2008–09 and 2009–10 school years there was "a huge decrease in allotment, both in staff allotment and actual dollars. In short, we didn't and couldn't pay as much attention to cohort 2 principals as we did to cohort 1."[7] This variation in year-to-year principal training and support likely shapes the empirical results that we present and discuss later in the paper.

4. DATA

Data for this paper consist of CPS administrative, personnel, and test score data from the 2005–06 school year to the 2010–11 school year. As the intervention occurred at the school level, student-level and teacher-level data records were aggregated up to the school level. Administrative data collected on students include basic demographic information, such as gender and race/ethnicity, as well as information on poverty level and students with special education needs. In this analysis, we use school-level characteristics such as student enrollment levels, the distribution of race/ethnicity, gender, students qualifying for free/reduced lunch, and special education students, which were generated from student-level CPS data files. These data allow us to check for covariate balance among the student characteristics across schools in the cohort 1 and

7. Sheila Cashman (EITP year 1 Project Coordinator and year 2 Project Manager), personal communication with authors, 17 June 2013.

cohort 2 groups, as well as to obtain more precise estimates of treatment effects. Further, the inclusion of student demographic characteristics enables us to identify heterogeneous treatment effects by school composition.

Teacher personnel data include teacher-level data about tenure status, years of experience in the district, demographic information, level of education attained, and certification status. Because this intervention was targeted at teachers, these variables are particularly important because they enable us to identify whether the pilot had different effects in schools with a greater proportion of tenured teachers or of teachers with higher levels of education, for example.

The outcome of interest in this paper is student achievement. Students in Illinois take the Illinois Standards Achievement Test (ISAT) in reading and mathematics in grades 3–8. The administration of the ISAT typically occurs in March of each school year. The ISAT is vertically scaled across grade levels, so students in upper grade levels will, on average, have higher scores than students in lower grade levels as an artifact of the scale score. In our analysis, we use a school-level measure that has been standardized across schools within the analytic sample. Because the test is vertically scaled, we first averaged student scores within grade level at a school. Then we created the school-level scale score by averaging the grade-level scale scores within a given school in order to account for any differences in grade sizes across the schools.[8] This paper uses ISAT data from six school years, 2005–06 to 2010–11 (three pre-policy and three post-policy years).

Table 4 presents mean pretreatment school-level characteristics for the cohort 1 and cohort 2 schools. Overall, there were forty-four schools that received the EITP treatment in the 2008–09 school year, and forty-eight cohort 2 schools that implemented the pilot in the following school year (2009–10). Between the cohort 1 and cohort 2 schools, the mean school characteristics appear to be quite balanced. The biggest difference between the two groups is the percentage of African American students, though this difference is not statistically significant. There are no statistically significant differences in the proportion of students who qualify for free/reduced-price lunch or who have been identified as having special education needs, nor are there differences between pre-treatment test scores in reading or math on the ISAT. This

8. Averaging at the grade level and then across grades within the school to generate the average scale test score in the school helps to alleviate the potential problem that may arise if a school has larger than expected enrollment in grade 4, for example. Because of the vertical scale of the ISAT, a school-level average that does not first account for the abnormally large fourth grade enrollment would be lower than average, but it would be artificially low. It is not necessarily the case that the school is underperforming but that the average test score in that school is low because they have a large number of fourth graders in their student population.

Table 4. Baseline Characteristics for Cohort 1 and Cohort 2 Elementary Schools

School Characteristic	Cohort 1 Mean (SD)	Cohort 2 Mean (SD)	*p*-Value of Difference
Enrollment	448.2 (209.8)	497.8 (296.0)	0.268
% Female	49 (2.7)	50 (3.1)	0.729
% African American	60 (40.1)	67 (37.4)	0.255
% Hispanic	24 (28.7)	20 (25.4)	0.511
% White	11 (17.5)	8 (13.1)	0.343
% Asian	5 (9.1)	5 (8.9)	0.916
Prop-IEP	0.13 (0.043)	0.14 (0.059)	0.455
Prop-FRPL	0.82 (0.232)	0.83 (0.207)	0.978
Math achievement	0.036 (1.083)	−0.032 (0.930)	0.826
Reading achievement	−0.008 (1.106)	0.007 (0.906)	0.829

Notes: Mean (standard deviation) of school characteristics for the 2008–09 school year. There are 44 cohort 1 schools and 49 cohort 2 schools in the sample. Math and Reading achievement are for the 2008 ISAT (given in the spring of the 2007–08 school year) and are standardized within sample. Prop-IEP: the proportion of students in a school in receipt of an individualized education plan; Prop-FRPL: the proportion of students receiving free or reduced-price lunch. Because elementary schools were randomly assigned to treatment within one of four local instructional areas, the *p*-value of difference of means is adjusted for area fixed effects. None of the reported coefficients is statistically different from zero at traditional levels of significance (e.g., $\alpha = .05$).

evidence suggests that the randomization of schools worked as intended. Further, regardless of the inclusion or exclusion of the school-level covariates in our analyses, the estimates of treatment effects remain stable.

Table 5 presents the teacher characteristics in the cohort 1 and cohort 2 schools. On average, in both cohort 1 and cohort 2 schools, 76 percent of teachers are tenured. Over half have master's degrees (on average, 59 percent in cohort 1 schools and 60 percent in cohort 2 schools), and very few have National Board Certification. Most of the teachers are women (on average, 85

Table 5. Teacher Characteristics for Cohort 1 and Cohort 2 Elementary Schools

Teacher Characteristic	Cohort 1 Mean (SD)	Cohort 2 Mean (SD)	*p*-Value of Difference
Number of teachers	28.4 (11.3)	29.9 (15.7)	0.451
% Female	85 (7.7)	85 (8.2)	0.988
% African American	38 (30.0)	42 (28.4)	0.385
% Hispanic	11 (15.5)	11 (12.5)	0.932
% White	46 (19.9)	42 (20.7)	0.301
% Asian	4 (5.7)	4 (5.1)	0.674
Age (years)	44.4 (4.18)	44.9 (4.09)	0.593
Experience (years)	11.8 (2.63)	12.1 (2.74)	0.602
Master's degree (%)	59 (12.4)	60 (12.2)	0.606
National Board Certification (%)	3 (4.2)	2 (3.5)	0.283
Tenured (%)	76 (11.9)	76 (15.1)	0.743

Notes: Mean (standard deviation) of teacher characteristics for the 2008–09 school year. There are 44 cohort 1 schools and 49 cohort 2 schools in the sample. Teacher tenure information is unavailable for two cohort 1 and four cohort 2 schools. Experience: the number of years a teacher has taught in CPS (not including any experience outside of the district). Because elementary schools were randomly assigned to treatment within one of four local instructional areas, the *p*-value of difference of means is adjusted for area fixed effects. None of the reported coefficients is statistically different from zero at traditional levels of significance (e.g., $\alpha = .05$).

percent in both groups of schools). The covariate balance across the two groups of schools is striking, which is especially important because this intervention was targeted at improving teacher practice. We can therefore attribute any differences in teacher performance at the end of 2008–09, after cohort 1 schools implement the Danielson pilot, to the intervention itself.

We also explored the extent of covariate balance on measures of school working conditions relevant to the teacher evaluation pilot, including school leadership, instructional quality, and school learning climate (see table 6). Using data from the spring 2007 administration of the CCSR biennial teacher

Table 6. Summary of CCSR Survey Measures of School Climate

CCSR Survey Measure	Description
	Leadership Measures
Principal instructional leadership (t)	The principal is an active and skilled instructional leader who sets high standards for teaching and student learning.
Teacher–principal trust (t)	Teachers and principals share a high level of mutual trust and respect.
	Instructional Quality Measures
Math instruction (s)	Students interact with course material and one another to build and apply knowledge in their math classes.
Quality of student discussion (t)	Students participate in classroom discussions that build their critical thinking skills.
Academic personalism (s)	Teachers connect with students in the classroom and support them in achieving academic goals.
Academic press (s)	Teachers expect students to do their best and to meet academic demands.
	Learning Climate Measures
Safety (s)	Students feel safe both in and around the school building, and while they travel to and from home.
Student–teacher trust (s)	Students and teachers share a high level of mutual trust and respect.
Peer support for academic work (s)	Students demonstrate behaviors that lead to academic achievement.

Notes: CCSR survey measures from the spring 2007 administration of the teacher and student surveys. The survey respondent (student [s] or teacher [t]) is indicated in parentheses next to the CCSR survey measure. Using Rasch analysis, CCSR creates measures of school climate from items on the teacher and student surveys. The CCSR measures are reliable and have been validated using school-level test score data. See http://ccsr.uchicago.edu/surveys for more information on the CCSR surveys.

and student surveys,[9] the most recent pre-treatment data available, we find no statistically significant differences in pre-treatment measures of school working conditions across the cohort 1 and cohort 2 schools.

5. EMPIRICAL STRATEGY

The key empirical issue in any observational study of a school-level intervention is the potential that factors related to a school's productivity are also related to its propensity to participate in the intervention itself. In the case of the EITP, higher-achieving schools (those that experience lower teacher turnover[10] and

9. CCSR surveys all principals, teachers, and six through twelfth grade students in the spring of every other school year (and every school year beginning in 2011) since 1995.
10. One potential concern along these lines is that teachers differentially sorted into different schools after they learned about the evaluation pilot. We do not believe this is an issue given the timing of teacher hiring in CPS combined with the timing of notification to schools about pilot

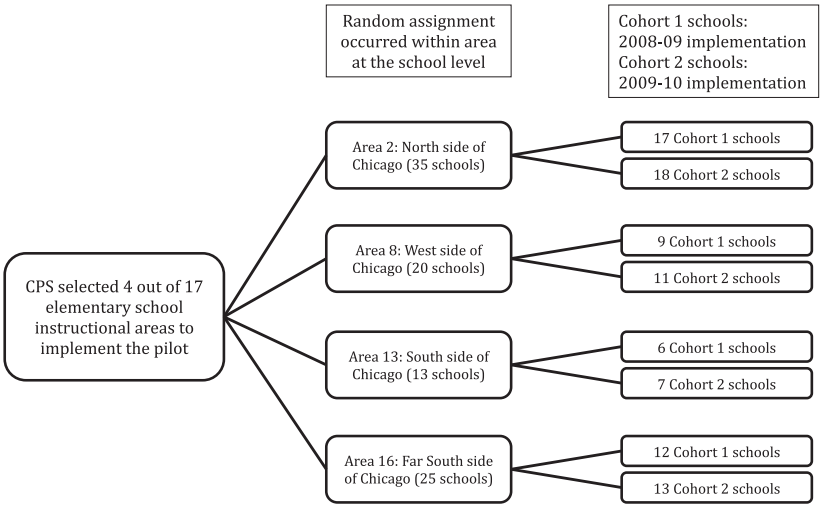

Figure 1. Experimental Design and Randomization Process.

have higher quality principals more capable of incorporating the additional responsibilities that an observation-based evaluation system requires) may be more willing to participate in a novel teacher evaluation system. In the presence of this type of nonrandom sorting, any observed differences in school achievement will likely be misattributed to the intervention itself. However, we are able to leverage the random assignment of schools to EITP in the first year of program implementation to consistently estimate the impact of the process-oriented teacher evaluation pilot on student test score performance.

We take advantage of a unique randomized control trial design. CPS, in partnership with CCSR, selected four elementary school instructional areas (of the seventeen elementary areas in the city at the time) that would implement the pilot program. These areas are located in different parts of the city, and they serve different populations of CPS students with varying needs. Within each of the four instructional areas, elementary schools were randomly selected to participate in the pilot. Schools with first-year principals and those slated for closure (at the end of the 2008–09 school year) were excluded from the randomization pool. Figure 1 illustrates the randomization process. Schools selected into the treatment group implemented the program in the 2008–09 school year (cohort 1). The cohort 2 schools implemented the pilot the following school year in 2009–10 (year 2). The randomization process resulted in

participation. In CPS, school teaching rosters for a given school year are generally solidified in the prior spring. Central office notified pilot principals during the summer of 2008. Teachers did not receive training on the pilot until the week before the first day of school in late August 2008. As such, we think it is unlikely that teachers would have sought new assignments in other schools or districts as a consequence of pilot participation in the first year.

forty-four cohort 1 schools and forty-nine cohort 2 schools in the 2008–09 school year.[11]

To estimate the impact of the teacher evaluation pilot on a school's math and reading achievement among our sample of CPS elementary schools, we estimate variants of the following basic model:

$$Y_i = \beta_0 + \beta_1(Pilot_i) + \boldsymbol{X_i}`\Gamma + \theta_g + \varepsilon_i, \quad (1)$$

where Y_i is a school achievement outcome for school i; $Pilot_i$ is an indicator variable that equals one if school i was randomly assigned to participate in the teaching pilot in the 2008–09 school year, and zero otherwise; $\mathbf{X}$ is a vector of school covariates, including student enrollment, the proportion of female students, the proportion of students by race/ethnicity, the proportion of special education students, and the proportion of students receiving free or reduced-price lunch; and ε_i is a random error term. Because the randomization was done at the instructional area level—schools were randomly assigned within one of the four elementary instructional areas—we also include area fixed effects (θ_g) to account for the block design of the experimental study. Among the forty-four elementary schools randomly assigned to the teaching pilot, there was full participation. Moreover, none of the forty-nine cohort 2 schools participated in the pilot in the initial year of implementation. As a result, $\widehat{\beta_1}$ estimates both the intent-to-treat as well as the treatment-on-the treated effects of the first year of the evaluation pilot.

6. FINDINGS

Graphical Evidence

Figures 2 and 3 capture the unadjusted trends in math and reading achievement, respectively, for the experimental sample and for all CPS elementary schools. In the pre-policy years (2005–06 through 2007–08), math and reading achievement trends in the experimental sample are very similar to district-wide trends, with evidence that reading achievement in the experimental sample increased at a slightly faster rate than the district as a whole for both cohort 1 and cohort 2 schools. At the end of the first post-policy year, cohort 1 schools' math performance appears to have grown at a faster rate between the 2007–08 and 2008–09 school years than both the cohort 2 schools as well as the district sample, and then increased at approximately the same rate as the cohort 2 (and all CPS elementary) schools in the second and third post-policy years. As shown in figure 3, reading achievement in the

11. At the end of the 2008–09 school year, one cohort 2 school closed, reducing the total EITP school sample from 93 to 92 schools. Our empirical findings are robust to the exclusion of this one school from the year 1 (2008–09) results.

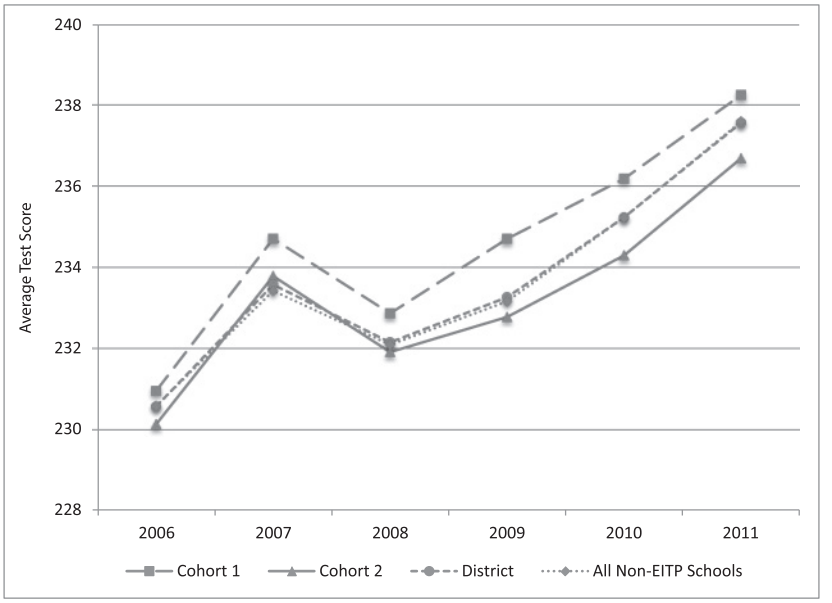

Figure 2. Unadjusted ISAT Math Trends.
Notes: Math achievement is shown in average scale scores from the spring administration of the Illinois Standards Achievement Test (ISAT). Each point represents the average ISAT score for schools in the denoted category. There are 44 cohort 1 schools and 49 cohort 2 schools (through 2009, after which there are 48 cohort 2 schools). The district trend includes 484 elementary schools through 2009 and 483 schools in 2010 and 2011, due to one cohort 2 school closing at the end of the 2008–09 school year. The non-EITP schools trend line includes 391 elementary schools that did not participate in the teacher evaluation program.

cohort 1 schools continued along its pre-policy trend at a constant rate. Reading achievement in the cohort 2 schools, however, declined by approximately 0.45 scale score points (or 0.04 standard deviations) at the end of the first post-policy year, while reading achievement increased by approximately 0.1 scale score points (or less than 0.01 standard deviations) in the all district and non-experimental schools. We are unable to reject the null that the cohort 2 schools' reading achievement trend between 2007–08 and 2008–09 is the same as the reading achievement trend for the district-wide sample or the all non-experimental school sample.[12] In subsequent analyses we show that cohort 2's deviation from its own pre-policy trend does not influence our main experimental results.

12. We conducted a *t*-test to assess the following null hypothesis: H_0: $\text{Achievement}_{\text{Cohort2 (08–09)}} = \text{Achievement}_{\text{District (08–09)}}$, where $\text{Achievement}_{\text{Cohort2 (08–09)}}$ and $\text{Achievement}_{\text{District (08–09)}}$ are the achievement trends (the difference in average achievement between the 2007–08 and 2008–09 school years) for the cohort 2 schools ($N = 49$) and district schools ($N = 484$), respectively. We are unable to reject the null hypothesis that the trends are the same (t-statistic = –.869, p-value = .386).

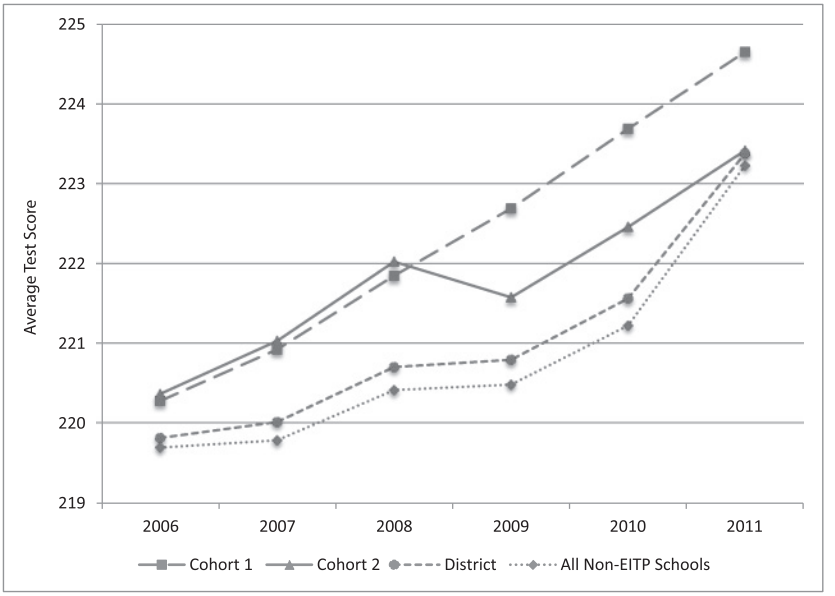

Figure 3. Unadjusted ISAT Reading Trends.
Notes: Reading achievement is shown in average scale scores from the spring administration of the Illinois Standards Achievement Test (ISAT). Each point represents the average ISAT score for schools in the denoted category. There are 44 cohort 1 schools and 49 cohort 2 schools (through 2009, after which there are 48 cohort 2 schools). The district trend includes 484 elementary schools through 2009 and 483 schools in 2010 and 2011, due to one cohort 2 school closing at the end of the 2008–09 school year. The non-EITP schools trend line includes 391 elementary schools that did not participate in the teacher evaluation program.

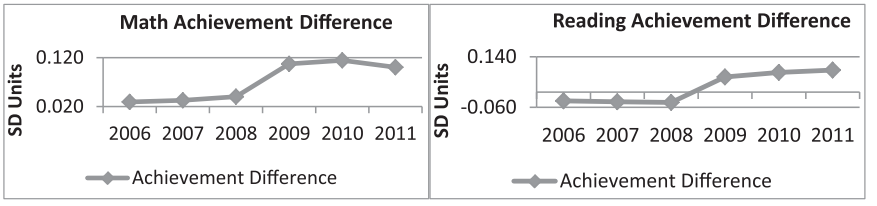

Figure 4. Achievement Differences between Cohort 1 and Cohort 2 Schools.
Notes: Each point represents the difference in average school achievement, in standard deviation units, between cohort 1 and cohort 2 schools. The average math and reading achievement are regression-adjusted averages net of area fixed effects, and standardized within sample (and within year) for cohort 1 and cohort 2 schools.

Experimental Estimates of EITP

We now consider the magnitude of the impact of the evaluation pilot on math and reading achievement to provide empirical estimates of the trends observed in figures 2 and 3. Figure 4 shows the differences in average achievement for the cohort 1 and cohort 2 schools. As illustrated in figure 4, we find no statistically significant differences between cohort 1 and cohort 2 schools in

pre-treatment (e.g., prior to the 2008–09 school year) average math or reading achievement, net of area fixed effects. This provides further evidence that the randomization of schools generated balance of covariates prior to implementation of EITP.

The experimental estimates confirm the achievement trends illustrated in figure 4 and provide additional support for both the short-term impact (i.e., the impact of the teacher evaluation pilot at the end of year 1, 2008–09) of the evaluation pilot as well as the persistent gap in achievement, even after the cohort 2 schools received the evaluation intervention. Table 7 summarizes the average achievement effects of the teacher evaluation pilot on a school's math achievement, and table 8 summarizes the impact on reading achievement. At the end of the first year of implementation (2008–09), the effect we find in math is 0.05 standard deviations, but it is not significantly different from zero (see column 4 of table 7). In years 2 and 3 post-policy (2009–10 and 2010–11), the gap in math achievement between cohort 1 and cohort 2 schools remains statistically indistinguishable from zero (see columns 8 and 12 in table 7).

At the end of 2008–09, we find statistically significant impacts on reading achievement that are double the size of the math effects. Specifically, net of school covariates and controlling for the schools' baseline reading achievement, reading achievement in cohort 1 schools improved by 0.10 standard deviations compared with cohort 2 schools (see column 4 of table 8).

The gap in reading achievement remains in year 2, and although the year 3 pilot school effect on reading is not statistically different from zero (because of the large standard error on the pilot coefficient), it is not statistically different from either the year 2 or year 3 effects (see columns 8 and 12 in table 8).

The school-level achievement gap provides one empirical benchmark for understanding the magnitude of the observed reading effect. According to Bloom et al. (2008), a 0.10 standard deviation effect size is equivalent to closing between one quarter to one half of the performance gap between weak schools (those at the 10th percentile of the achievement distribution) and average schools (those at the 50th percentile) in large urban districts.

Although we observe a statistically significant jump in achievement for the cohort 1 schools relative to the cohort 2 schools after the first year of the pilot, the difference in reading achievement remained but did not continue to grow in subsequent school years, even in light of the cohort 2 schools participating in the evaluation system beginning in the 2009–10 school year. The fact that cohort 2 schools never saw the benefit that the initial implementers did is likely explained by changes in district leadership and reduced support for principals implementing the pilot system. Differences in training for the cohort 1 and cohort 2 school principals are shown in table 3. We discuss in greater detail

Table 7. Impact of Teaching Evaluation Pilot on Math Achievement

	Year 1				Year 2				Year 3			
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**	**(10)**	**(11)**	**(12)**
Pilot	0.136	0.107	0.127	0.054	0.128	0.114	0.146	0.080	0.110	0.100	0.120	0.066
	(0.2109)	(0.1826)	(0.1310)	(0.0542)	(0.2110)	(0.1832)	(0.1380)	(0.0816)	(0.2111)	(0.1836)	(0.1368)	(0.0894)
School characteristics			X	X			X	X			X	X
Baseline math achievement				X				X				X
Area fixed effects		X	X	X		X	X	X		X	X	X
No. of schools	93	93	93	93	92	92	92	92	92	92	92	92
R^2	0.0046	0.2669	0.6942	0.9463	0.0041	0.2689	0.7030	0.9100	0.0030	0.2652	0.6550	0.8450

Notes: Coefficients (with robust standard errors) reported are in standard deviation units and represent the intent-to-treat effect of the teacher evaluation Pilot on math achievement. Year 1 effects for the 2008–09 school year; Year 2 effects for the 2009–10 school year; and Year 3 effects for the 2010–11 school year. School characteristics include: enrollment, gender, race/ethnicity, the proportion of special education students, and the proportion of students receiving free or reduced-price lunch. Baseline math achievement: a school's pre-treatment math achievement (standardized) for the 2007–08 school year. None of the reported coefficients is statistically different from zero at traditional levels of significance (e.g., $\alpha = .05$).

Table 8. Impact of Teaching Evaluation Pilot on Reading Achievement

	Year 1				Year 2				Year 3			
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**	**(10)**	**(11)**	**(12)**
Pilot	0.087	0.060	0.107	0.099**	0.089	0.078	0.104	0.115*	0.094	0.087	0.091	0.1200
	(0.2116)	(0.1871)	(0.1246)	(0.0463)	(0.2119)	(0.1853)	(0.1257)	(0.0685)	(0.2116)	(0.1813)	(0.1262)	(0.0803)
School characteristics			X	X			X	X			X	X
Baseline reading achievement				X				X				X
Area fixed effects		X	X	X		X	X	X		X	X	X
No. of schools	93	93	93	93	92	92	92	92	92	92	92	92
R^2	0.0019	0.2335	0.7226	0.9620	0.0020	0.2581	0.7466	0.9292	0.0022	0.2894	0.6971	0.8803

Notes: Coefficients (with robust standard errors) reported are in standard deviation units and represent the intent-to-treat effect of the teacher evaluation Pilot on reading achievement. Year 1 effects for the 2008–09 school year; Year 2 effects for the 2009–10 school year; and Year 3 effects for the 2010–11 school year. School characteristics include: enrollment, gender, race/ethnicity, the proportion of special education students, and the proportion of students receiving free or reduced-price lunch. Baseline reading achievement: a school's pre-treatment reading achievement (standardized) for the 2007–08 school year.

*Coefficients statistically significant at the 10 percent level; **significant at the 5 percent level.

Table 9. Heterogeneous Treatment Effects: School Achievement

	Math			Reading		
	Year 1	**Year 2**	**Year 3**	**Year 1**	**Year 2**	**Year 3**
Pilot	.061	.0808	.0656	.107**	.116*	.122
	(.0530)	(.0812)	(.0894)	(.0441)	(.0682)	(.0811)
Baseline achievement	.847***	.833***	.767***	.854***	.806***	.775***
	(.0574)	(.0759)	(.0859)	(.0566)	(.0540)	(.0671)
Pilot * Baseline achievement	.078	.0178	−.006	.087*	.0319	.023
	(.0488)	(.0615)	(.0859)	(.0476)	(.0527)	(.0784)
School characteristics	X	X	X	X	X	X
Area fixed effects	X	X	X	X	X	X
No. of schools	93	92	92	93	92	92
R^2	.9476	.9101	.8450	.9636	.9294	.8805

Notes: Coefficients (with robust standard errors) reported are in standard deviation units. Year 1 effects for the 2008–09 school year; Year 2 effects for the 2009–10 school year; and Year 3 effects for the 2010–11 school year. School characteristics include: enrollment, gender, race/ethnicity, the proportion of special education students, and the proportion of students receiving free or reduced-price lunch. Baseline achievement: a school's pre-treatment math or reading achievement (standardized) for the 2007–08 school year.

*Coefficients statistically significant at the 10 percent level; **significant at the 5 percent level; ***significant at the 1 percent level.

the potential reasons for this persistent gap in school achievement later in the paper.

Heterogeneous Effects by School Composition

Although there is clear evidence of a positive impact of the evaluation pilot on reading achievement after the first year of implementation, there may also be significant heterogeneity in the impact of the pilot. To explore this heterogeneity, we ask whether, and to what extent, the evaluation pilot differentially impacted achievement in cohort 1 schools with different characteristics. We consider the impact of the evaluation pilot in schools with different levels of achievement, schools that serve students from different economic circumstances, schools with different racial and ethnic profiles, and schools serving more students with individual learning needs. Tables 9–12 summarize the heterogeneous effects of the evaluation pilot.

With respect to a school's reading achievement, higher-achieving schools realized a bigger benefit of the evaluation pilot than lower-achieving schools (see table 9 and figure 5). After the first year of the pilot, higher-achieving schools (those that are one standard deviation above the mean in prior reading achievement) improved by 0.09 standard deviations over lower-performing pilot schools. This coefficient is approximately the same size as the main

Table 10. Heterogeneous Treatment Effects: School Poverty

	Math			Reading		
	Year 1	**Year 2**	**Year 3**	**Year 1**	**Year 2**	**Year 3**
Pilot	0.068	0.090	0.080	0.114**	0.125*	0.129
	(0.0525)	(0.0777)	(0.0650)	(0.0450)	(0.0648)	(0.0803)
% FRPL	−0.004	−0.018	0.105	−0.027	−0.045	0.0635
	(0.0466)	(0.0721)	(0.1173)	(0.0490)	(0.0621)	(0.0932)
Pilot*% FRPL	−0.091**	−0.111*	−0.156*	−0.099**	−0.129**	0.101
	(0.0415)	(0.0567)	(0.0939)	(0.0474)	(0.0531)	(0.0892)
School characteristics	X	X	X	X	X	X
Area fixed effects	X	X	X	X	X	X
No. of schools	93	92	92	93	92	92
R^2	0.9478	0.9125	0.8498	0.9637	0.9325	0.8823

Notes: Coefficients (with robust standard errors) reported are in standard deviation units. Year 1 effects for the 2008–09 school year; Year 2 effects for the 2009–10 school year; and Year 3 effects for the 2010–11 school year. FRPL: the standardized proportion of students in a school receiving federally subsidized free or reduced-price lunch. For the 2008–09 school year, the mean (SD) proportion of students in receipt of free or reduced-price lunch within sample of cohort 1 and cohort 2 schools is 0.826 (0.218); for the 2009–10 school year, the mean (SD) is 0.854 (0.216); and for the 2010–11 school year, the mean (SD) is 0.833 (0.229). School characteristics include: enrollment, gender, race/ethnicity, baseline achievement (the pre-treatment math or reading achievement (standardized) for the 2007–08 school year), and the proportion of special education students.
*Coefficients statistically significant at the 10 percent level; **significant at the 5 percent level.

effect of the pilot, which negates any positive impact that the pilot may have had in schools that are less than one standard deviation below the mean test score and doubles the impact for schools that are one standard deviation above the mean. We do not find any additional benefit to high-achieving pilot schools in the second and third post-policy years. Although we are unable to reject zero impacts of the pilot after the first year on math achievement among high achieving pilot schools, the magnitude of the coefficient (0.08 standard deviations) is not statistically different from the effect on reading achievement (see table 9 and figure 6).

Low-poverty schools benefited more from the evaluation pilot than high-poverty schools (see table 10 and figures 7 and 8). Indeed, at the end of the first post-policy year, lower-poverty schools (those that are one standard deviation below the mean share of students receiving free or reduced-price lunch) improved by 0.09 standard deviations in math and 0.10 standard deviations in reading beyond the pilot effect at schools with average poverty level. The differential effect of the pilot persisted even after the first year of the evaluation pilot.

Table 11. Heterogeneous Treatment Effects: Race

	Math			Reading		
	Year 1	**Year 2**	**Year 3**	**Year 1**	**Year 2**	**Year 3**
Pilot	0.066	0.087	0.068	0.112**	0.131*	0.133*
	(0.0532)	(0.0784)	(0.0849)	(0.0460)	(0.0659)	(0.0777)
% Minority	−0.139	0.309	0.197	1.57	2.62	0.131
	(1.664)	(1.873)	(0.7673)	(1.384)	(1.925)	(0.6851)
Pilot * % Minority	−0.062	−0.061	−0.105	−0.048	−0.067	−0.088
	(0.0427)	(0.0646)	(0.0826)	(0.0435)	(0.0598)	(0.0779)
School characteristics	X	X	X	X	X	X
Area fixed effects	X	X	X	X	X	X
No. of schools	93	92	92	93	92	92
R^2	0.9466	0.9106	0.84702	0.9617	0.9288	0.8816

Notes: Coefficients (with robust standard errors) reported are in standard deviation units. Year 1 effects for the 2008–09 school year; Year 2 effects for the 2009–10 school year; and Year 3 effects for the 2010–11 school year. *Minority%:* the standardized proportion of students in a school that are African-American or Hispanic. For the 2008–09 school year, the mean (SD) proportion of minority students within sample of cohort 1 and cohort 2 schools is .853 (.210); for the 2009–10 school year, the mean (SD) is 0.848 (0.218); and for the 2010–11 school year, the mean (SD) is 0.841 (0.220). School characteristics include: enrollment, gender, race/ethnicity, baseline achievement (the pre-treatment math or reading achievement (standardized) for the 2007–08 school year), the proportion of special education students, and the proportion of students receiving free or reduced price lunch.
*Coefficients statistically significant at the 10 percent level; **significant at the 5 percent level.

Although there is heterogeneity in the impact of the teacher evaluation pilot as a function of both pre-treatment achievement and school poverty, there is no evidence that the effect of the pilot differentially affected schools that serve different shares of minority students and special education students (those that receive an individualized education plan). Tables 11 and 12 summarize these results. In results not presented here (but available from the authors upon request), we also explored whether the pilot differentially affected schools with teachers who differed on years of experience, the proportion of master's degree holders, and the proportion of tenured teachers. We did not find any evidence that pilot schools with different teacher compositions realized different impacts on either reading or math achievement.

Interrupted Time Series Estimates

As previously discussed, the cohort 2 schools' reading achievement declined between the 2008 and 2009 school years (see figure 3), a deviation from its own pre-policy trend, and reading achievement remained approximately unchanged for all district schools and all non-EITP schools (and continued to improve for the cohort 1 schools). Although the cohort 2 schools' trend does

Table 12. Heterogeneous Treatment Effects: Individualized Education Program

	Math			Reading		
	Year 1	**Year 2**	**Year 3**	**Year 1**	**Year 2**	**Year 3**
Pilot	0.053	0.081	0.066	0.098**	0.116*	0.120
	(0.0553)	(0.0819)	(0.0898)	(0.0469)	(0.0689)	(0.0807)
%IEP	−0.004	−0.016	−0.076	0.001	−0.001	−0.009
	(0.0348)	(0.0539)	(0.0628)	(0.0301)	(0.0535)	(0.0529)
Pilot * %IEP	−0.005	0.024	−0.008	−0.005	0.021	−0.013
	(0.0498)	(0.0622)	(0.0787)	(0.0403)	(0.0518)	(0.0725)
School characteristics	X	X	X	X	X	X
Area fixed effects	X	X	X	X	X	X
No. of schools	93	92	92	93	92	92
R^2	0.9463	0.9102	0.8451	0.9620	0.9293	0.8804

Notes: Coefficients (with robust standard errors) reported are in standard deviation units. Year 1 effects for the 2008–09 school year; Year 2 effects for the 2009–10 school year; and Year 3 effects for the 2010–11 school year. *IEP* is the standardized proportion of students in a school receiving individualized education services. For the 2008–09 school year, the mean (sd) proportion of students in receiving individualized education services within sample of cohort 1 and cohort 2 schools is 0.133 (0.052); for the 2009–10 school year, the mean (sd) is 0.135 (0.049); and for the 2010–11 school year, the mean (sd) is 0.140 (0.049). School characteristics include: enrollment, gender, race/ethnicity, baseline achievement (the pre-treatment math or reading achievement (standardized) for the 2007–08 school year), and the proportion of students receiving free or reduced price lunch.
*Coefficients statistically significant at the 10% level; **significant at the 5% percent level.

not differ in a statistically significant way from district-wide trends, the deviation from its own pre-policy trends raises some concerns about the role of the cohort 2 schools as the counterfactual in estimating the impact of the teacher evaluation pilot. It does appear, however, that the cohort 2 schools were picking up a district-wide reading performance trend; notably, stagnant reading achievement growth between 2008 and 2009. Moreover, because the cohort 1 schools' reading achievement trend maintained a constant positive slope over the study period, it is likely that the teacher evaluation policy insulated the cohort 1 schools from this district-wide performance stagnation that would have affected the cohort 1 schools in the absence of the teacher evaluation policy, and therefore allowed the cohort 1 schools to continue along their pre-policy trend.

To further explore the robustness of the main result—the improvement in reading achievement for the early adopters (i.e., the cohort 1 schools) at the end of the first year of the teacher evaluation pilot—we consider a different set of comparison schools: all non-EITP elementary schools. We implement a comparative interrupted time series (CITS) approach, which allows us to compare the deviation from pre-policy achievement trends among the cohort

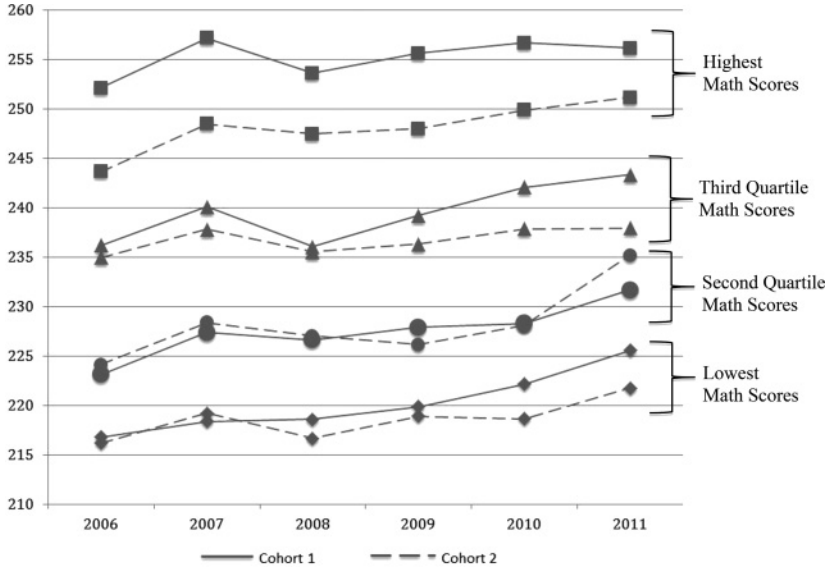

Figure 5. Math Test Score Trends by Pre-Treatment Math ISAT Quartile.
Notes: Math achievement is shown in average scale scores from the spring administration of the Illinois Standards Achievement Test (ISAT). Each point represents the average ISAT score for schools in the denoted category. To place schools into math achievement quartiles, schools were ranked by pre-pilot math ISAT scores from spring 2008 and then divided into four groups. Sample sizes are the following: 10 cohort 1 schools and 13 cohort 2 schools in the top quartile with the highest reading scores; 9 cohort 1 schools and 14 cohort 2 schools in the third quartile; 15 cohort 1 schools and 8 cohort 2 schools in the second quartile; and 10 cohort 1 schools and 14 cohort 2 schools in the bottom quartile with the lowest math scores (after 2009, one cohort 2 school closed, reducing the sample of cohort 2 schools in the bottom quartile to 13 schools).

1 schools (the early adopters first implementing EITP in 2008–09) and cohort 2 schools (the late adopters first implementing EITP in 2009–10) to all district schools that never participated in the teacher evaluation pilot.[13] The CITS findings support the experimental evidence presented previously.

Following Dee and Jacob (2011), we estimate variants of the following regression model in the context of the CITS design:

$$\begin{aligned} Y_{st} = {} & \beta_0 + \beta_1 Year_t + \beta_2 EITP_t + \beta_3(years\_since\_EITP_t) + \beta_4(T_s * Year_t) \\ & + \beta_5(T_s * EITP_t) + \beta_6(T_s * years\_since\_EITP_t) + \beta_7(C_s * Year_t) \\ & + \beta_8(C_s * EITP_t) + \beta_9(C_s * years\_since\_EITP_t) + \beta_{10}\mathbf{X}_{st} + \theta_s + \varepsilon_{st}, \end{aligned} \tag{2}$$

where Y_{st} is achievement on the ISAT math or reading exam for school s in year t, $Year_t$ is a trend variable defined as $Year_t - 2005$ and starts at a value of 1 in

13. For the time series models, we use all elementary schools with complete data for the six-year time period, 2005–06 through 2010–11.

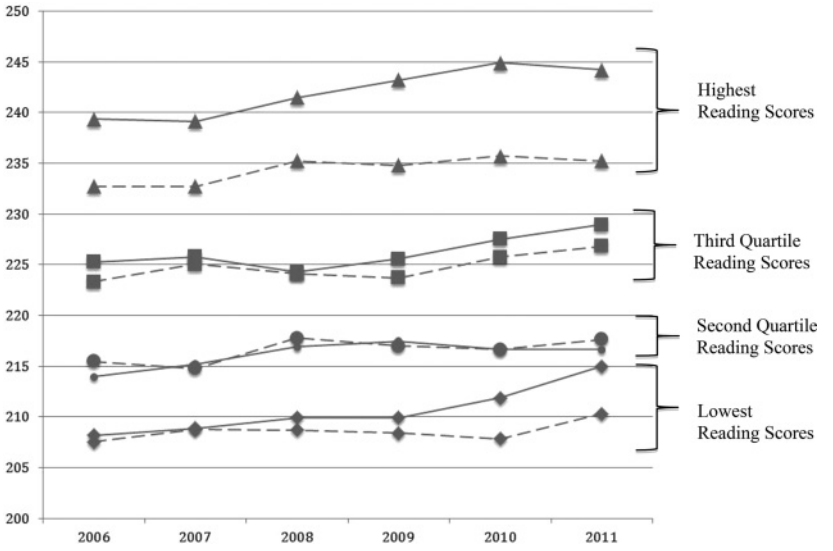

Figure 6. Reading Test Score Trends by Pre-Treatment Reading ISAT Quartile.
Notes: Reading achievement is shown in average scale scores from the spring administration of the Illinois Standards Achievement Test (ISAT). Each point represents the average ISAT score for schools in the denoted category. To place schools into reading achievement quartiles, schools were ranked by pre-pilot reading ISAT scores from spring 2008 and then divided into four groups. Sample sizes are the following: 9 cohort 1 schools and 14 cohort 2 schools in the top quartile with the highest reading scores; 10 cohort 1 schools and 13 cohort 2 schools in the third quartile; 14 cohort 1 schools and 9 cohort 2 schools in the second quartile; and 11 cohort 1 schools and 13 cohort 2 schools in the bottom quartile with the lowest reading scores (after 2009, one cohort 2 school closed, reducing the sample of cohort 2 schools in the bottom quartile to 12 schools).

the first year of the sample (2005–06 school year). $EITP_t$ is a dummy variable indicating the post-policy period, such that observations in the pre-policy period (2005–06 through 2007–08) take on a value of zero and observations during the teacher evaluation pilot period (2008–09 through 2010–11) take on a value of 1. The variable $years\_since\_EITP_t$ captures the number of years since a school first participated in the teacher evaluation pilot. For the cohort 1 schools, this variable takes on a value of 1 for the 2008–09 school year; for the cohort 2 schools first participating in the teacher evaluation pilot during the 2009–10 school year, this variable takes on a value of 1 for the 2009–10 school year. X_{st} captures time-varying school covariates, including enrollment, and the distribution of students by race, gender, and special education status. The variables θ_s and ε_{st} represent school fixed-effects and a mean-zero random error, respectively.

The variable T_s is a time-invariant variable that indicates whether a school was in the cohort 1 group of the teacher evaluation pilot (the treatment group schools are the forty-four early adopters, first implementing the teacher evaluation pilot in the 2008–09 school year). The variable C_s is a time-invariant

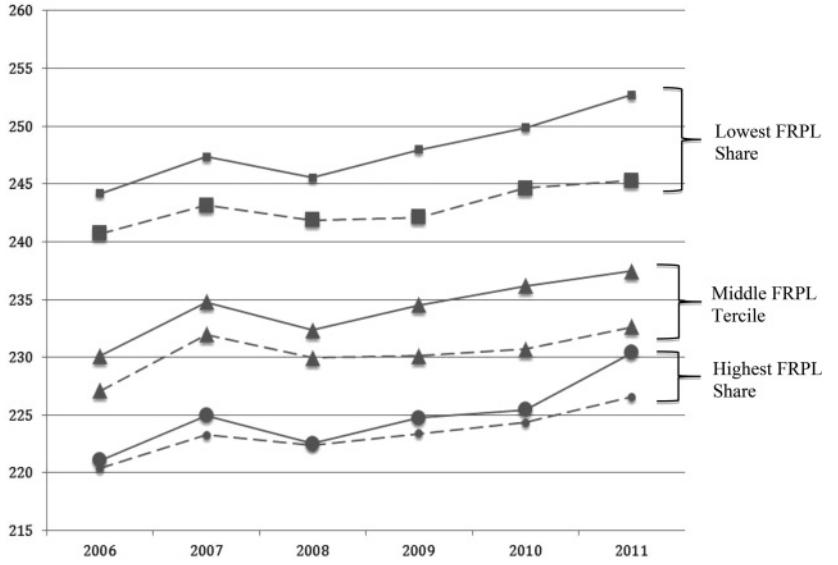

Figure 7. Math Test Score Trends by Pre-Treatment Free/Reduced-Price Lunch Tercile.
Notes: Math achievement is shown in average scale scores from the spring administration of the Illinois Standards Achievement Test (ISAT). Each point represents the average ISAT score for schools in the denoted category. To place schools into free/reduced-price lunch (FRPL) terciles, schools were ranked by pre-pilot free/reduced-price lunch rates from fall 2008 and then divided into four groups. Sample sizes are the following: 13 cohort 1 schools and 18 cohort 2 schools in the tercile with the lowest share of students eligible for FRPL; 16 cohort 1 schools and 15 cohort 2 schools in the middle tercile; and 15 cohort 1 schools and 16 cohort 2 schools in the tercile with the highest share of FRPL eligible students (after 2009, one cohort 2 school closed, reducing the sample of cohort 2 schools with the highest share of FRPL eligible students to 15 schools).

variable that indicates whether a school was in the cohort 2 group of the teacher evaluation pilot (the forty-nine late adopters, of which forty-eight schools first implemented the teacher evaluation pilot in the 2009–10 school year).

This regression specification allows us to estimate the impact of the teacher evaluation pilot for both the early and late adopter schools, relative to all non-EITP elementary schools (of which there are 391 in the sample). This impact is reflected in both a level shift in reading and math achievement (captured by β_5 and β_8 for the cohort 1 and cohort 2 schools, respectively) as well as a shift in the achievement trend (captured by β_6 and β_9 for the cohort 1 and cohort 2 schools, respectively). Therefore, the total estimated effect of the teacher evaluation policy on school achievement for cohort 1 schools at the end of the three-year post-policy period will be $\hat{\beta}_5 + 3(\hat{\beta}_6)$, and the total estimated effect for cohort 2 schools at the end of the post-policy period will be $\hat{\beta}_8 + 2(\hat{\beta}_9)$. The results are summarized in table 13.

We find that the impact of the teacher evaluation pilot on the reading and math achievement of early adopter schools (relative to all non-EITP schools) is nearly identical in magnitude to the experimental estimates. Specifically,

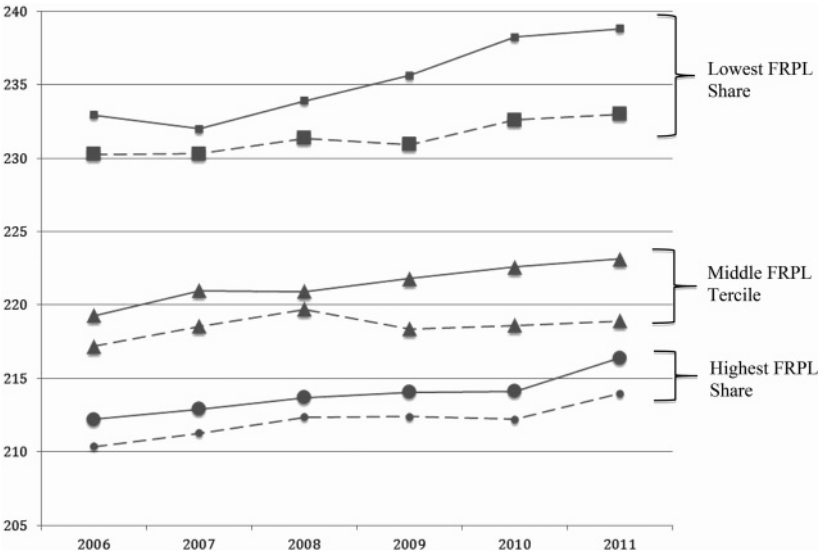

Figure 8. Reading Test Score Trends by Pre-Treatment Free/Reduced-Price Lunch Tercile.
Notes: Reading achievement is shown in average scale scores from the spring administration of the Illinois Standards Achievement Test (ISAT). Each point represents the average ISAT score for schools in the denoted category. To place schools into free/reduced-price lunch (FRPL) terciles, schools were ranked by pre-pilot free/reduced-price lunch rates from fall 2008 and then divided into four groups. Sample sizes are the following: 13 cohort 1 schools and 18 cohort 2 schools in the tercile with the lowest share of students eligible for FRPL; 16 cohort 1 schools and 15 cohort 2 schools in the middle tercile; and 15 cohort 1 schools and 16 cohort 2 schools in the tercile with the highest share of FRPL eligible students (after 2009, one cohort 2 school closed, reducing the sample of cohort 2 schools with the highest share of FRPL eligible students to 15 schools).

the estimate of β_5 picks up the shift in school achievement as a function of the teacher evaluation pilot, on the order of 0.10 standard deviations in reading and a (nonsignificant) 0.05 standard deviations in math.[14] Although we observe a shift in average achievement in the post-policy period for the cohort 1 schools, there is no evidence that the achievement trend (both for math and reading) is statistically different from zero, even as there appears to have been a reduction in the net difference in reading achievement between cohort 1 and non-EITP schools (see figure 3) by the end of the 2010–11 school year. For the late adopters, there is no statistically significant evidence that the cohort 2 schools realized an achievement benefit, relative to non-EITP

14. The magnitude of the intercept shift in reading achievement for the cohort 1 schools at the end of the first year (2008–09) provides additional evidence that the reading trend for the cohort 2 and non-EITP schools is not statistically different (both trends stagnate between 2008 and 2009). In particular, the magnitude of the achievement effect for cohort 1 schools relative to non-EITP schools is approximately the same (0.10 standard deviations) as the one-year effect relative to cohort 2 schools (see table 8). If the cohort 2 schools' achievement trend declined more than the non-EITP schools between 2008 and 2009, we would expect to see a smaller effect for the cohort 1 schools when we change the comparison group to all non-EITP schools. This is not what we observe at the end of the first post-policy year.

Table 13. Time Series Estimates

	(1) Math	(2) Math	(3) Math	(4) Reading	(5) Reading	(6) Reading
$T_s * EITP_t$	0.053	0.059	0.053	0.099*	0.103*	0.094*
	(0.0614)	(0.0518)	(0.0510)	(0.0588)	(0.0528)	(0.0521)
$T_s * (years\_since\_EITP)_t$	−0.052	−0.042	−0.036	−0.069	−0.064	−0.058
	(0.0791)	(0.0503)	(0.0501)	(0.0798)	(0.0550)	(0.0542)
$C_s * EITP_t$	−0.026	−0.054	−0.046	−0.077	−0.075	−0.068
	(0.1010)	(0.0498)	(0.0488)	(0.100)	(0.0482)	(0.0477)
$C_s * (years\_since\_EITP)_t$	−0.025	−0.041	−0.034	−0.094	−0.093**	−0.086**
	(0.0633)	(0.0384)	(0.0389)	(0.0632)	(0.0377)	(0.0384)
No. of schools	483	483	483	483	483	483
Sample size	2,898	2,898	2,898	2,898	2,898	2,898
School fixed effects		X	X		X	X
School covariates			X			X

Notes: Each column represents a separate regression. Coefficients are in standard deviation units and robust standard errors (clustered at the school level) are in parentheses. There are 44 schools that participated in the EITP pilot for the first time in the 2008–09 school year, and 48 schools that first participated in the EITP pilot in the 2009–10 school year. There are 391 non-EITP elementary schools in the sample. School characteristics include: enrollment, gender, race/ethnicity, and the proportion of special education students.
*Coefficients statistically significant at the 10 percent level; **significant at the 5 percent level.

schools, in either math or reading. Nevertheless, there is evidence that the reading achievement slope for cohort 2 schools declined relative to non-EITP schools—on the order of 0.09 standard deviations per year. The decline in the reading achievement trend for the cohort 2 schools relative to non-EITP schools can also be seen in figure 3, where the average reading achievement levels for the two sets of schools converge by the end of the 2011–12 school year.

7. DISCUSSION

The EITP represented a dramatic departure from the status quo teacher evaluation system in Chicago; its efficacy depended on a number of factors. These included the principals' capacity to provide targeted instructional guidance, their teachers' ability to respond to the instructional feedback in a manner that generated improvements in student achievement, and the extent of district-level support and training for principals who were primarily responsible for implementing a new teacher evaluation system. As a result, EITP was particularly human-capital-intensive, and relied on the human capital that already existed in the school (both principals and teachers) to generate improvements in school performance.

The pilot forced principals to make significant changes to how they conducted classroom observations and conferences with teachers, as well as to how they conceptualized teacher evaluation more generally. The intervention itself was time-intensive for the principals, who were required to participate in extensive training pre-intervention. Further, the newly implemented pilot system demanded enormous effort to evaluate teachers. Principals had to rate teachers on the new evaluation framework, and also to work with teachers in pre- and post-observation conferences to develop strategies to improve their instructional practice. Indeed, the principals' role under this new pilot system evolved from pure evaluation to one where the principal incorporated instructional coaching into a dual role as evaluator and formative assessor of a teacher's instructional practice. Certainly, more able principals can accomplish this new role more effectively than less able principals, and prior research found there was variation in principals' instructional coaching capacity (Sartain et al. 2011). A very similar argument can be made for the human capital demands that the new evaluation pilot placed on teachers. Indeed, higher human capital teachers are likely more able to incorporate principal feedback and assessment into their instructional practice.

As mentioned earlier, our results indicate that although the pilot evaluation system led to large short-term, positive effects on school reading performance, these effects were largely driven by schools that, on average, served higher-achieving and less-disadvantaged students. Indeed, for high-poverty schools—those that are one standard deviation above the mean level of poverty as measured by the share of students in receipt of free or reduced-price lunch—the net effect of the pilot after the first year is effectively zero. Why might we observe such heterogeneity in the impact of the evaluation pilot by school achievement and poverty?

If principal and teacher human capital is distributed unequally across schools—for example, if higher quality principals and teachers systematically sort into higher-achieving, lower-poverty schools—it should not be surprising that the impact of the pilot is also distributed unequally. Evidence suggests this very sorting occurs both in Chicago (Allensworth, Ponisciak, and Mazzeo 2009) and elsewhere (Ingersoll 2001; Hanushek, Kain, and Rivkin 2004; Boyd et al. 2009). As such, systems designed to improve teacher instructional practice should attend to the idiosyncratic context in which teachers do their work. Specifically, less-advantaged schools with, on average, lower performing teachers and harder-to-serve student populations, may require additional supports for human capital intensive interventions, such as observation-based evaluation systems, to generate improvements in student learning similar to more advantaged schools. This is relevant given evidence that students of

less-experienced teachers realize larger achievement gains when their teachers have higher performing colleagues (Jackson and Bruegmann 2009).

In addition to the human capital demands on principals and teachers, the nature of school-level implementation is critically important for the success of any new educational intervention. As previously discussed, the extent of principal training and district-level support varied dramatically across the two cohorts of schools—the early adopters in 2008 and the late adopters in 2009. We find that the cohort 1 schools experienced a significant achievement benefit after the first year of EITP participation, irrespective of whether we compare these schools to cohort 2 schools or to all other elementary schools that never participated in EITP. However, there is evidence that reading achievement in cohort 2 schools lagged behind the non-EITP schools after they first implemented the evaluation system. What role could district support play in generating these results?

Leadership turnover in CPS led to a decline in institutional and district support for EITP between the 2008–09 and 2009–10 school years. When the pilot program started in Chicago in 2008, few people were paying attention to teacher evaluation issues. Through its two years of planning work with the teachers' union, the district leadership demonstrated its commitment to the Danielson pilot and to evaluating teachers in a way that was systematic and fair. When introducing the pilot program for the first time to principals, the Chief Education Officer, Barbara Eason-Watkins, herself a former principal, personally delivered the message that the EITP pilot would be the district's cornerstone in improving the quality of teaching and instruction and increasing student learning.

Not long into the pilot's first year of implementation, however, CEO Arne Duncan left CPS (in early 2009) to serve as United States Secretary of Education. Duncan's arrival in Washington was followed by a national emphasis on refining teacher evaluation systems, but his departure from Chicago marked a move away from the rigorous year 1 implementation of the EITP pilot. Specifically, the district was under a new administration that de-emphasized the newly implemented teacher evaluation pilot and instead focused on performance monitoring, data usage, and accountability. The increased focus on accountability, a large, well-publicized budget deficit, and rumors of layoffs, led to a sense of insecurity among principals and teachers. Although the Danielson pilot was scaled up in year 2 (2009–10), doubling the number of schools implementing the pilot, district program staff reported that the budget did not increase.[15] This effectively limited the amount of support central

15. Sheila Cashman, personal communication with authors, 17 June 2013.

office could provide to principals, which in turn weakened the intervention and, we suspect, reduced the fidelity with which the pilot was implemented in schools. Indeed, CPS central office staff responsible for EITP oversight and school-level implementation indicated that, between the 2008–09 and 2009–10 school years, there was a significant decrease in both CPS staff and budgetary resources dedicated to cohort 2 principals in comparison with the level of support cohort 1 principals received during their first year of program participation. As a result, cohort 2 principals received fewer hours of training as well as different types of training than cohort 1 principals did in their first year of implementation (see table 3). Finally, in the summer of 2010, prior to the third year of implementation, CPS ended EITP. Just before this announcement, half of the principals in the district were set to receive Danielson Framework training, but the district canceled this training. As a result, there is little evidence that the Danielson Framework was used in any systematic way in year 3.[16] Our results are consistent with strong implementation in year 1 and weak implementation in subsequent years.

Finally, we consider our results in the Chicago context relative to Cincinnati's teacher evaluation program. Under Cincinnati's Teacher Evaluation System, launched during the 2000–01 school year, teachers were observed four times in a year, with no less than three of the four annual observations conducted by a trained evaluator external to the school (and only once by either the principal or another school administrator).[17] Like Cincinnati, the teacher evaluation system in CPS was based entirely on the Danielson (1996) classroom-observation protocol. A notable difference between Cincinnati and Chicago, however, is the nature of the evaluation system as an intervention. As discussed, Chicago's system is most accurately considered a school-level intervention, whereas Cincinnati's system may be considered a teacher-level intervention. Specifically, the annual evaluation of Cincinnati teachers depended exclusively on their year of hire, whereas all teachers in participating Chicago schools were evaluated twice annually, irrespective of date of hire.[18] In Cincinnati, the evaluation system relied on the quality of the external evaluators, who were primarily responsible for observing and evaluating teachers.

16. Interestingly, the debate about what classroom observation rubric to use in CPS has come full circle. In January 2010, the Illinois state legislature passed the Performance Evaluation Reform Act (PERA) in order to make Illinois more competitive in the RTTT competition. PERA required that indicators of student growth be a "significant factor" in teacher evaluation. For CPS, this legislation meant that a new teacher evaluation system had to be developed, which was to be used for the first time in the 2012–13 school year. That evaluation system (called REACH) uses the Danielson Framework for the classroom observation component.
17. Taylor and Tyler (2012) noted that each teacher was evaluated approximately once every five years.
18. In Cincinnati, only teachers hired during the 1999–2000 school year, for example, were observed and evaluated during the 2006–07 school year, with an average tenure of eight years at the time of evaluation (Taylor and Tyler 2012)

In Chicago, school principals were exclusively responsible for observing and evaluating all of their teachers, in addition to their normal duties as school leaders. Therefore, the quality of Chicago's system depended largely on principal capacity and institutional, district-level supports. Finally, where Taylor and Tyler (2012) used a teacher fixed effects approach, exploiting within-teacher variation over time in their participation in the evaluation system, in this paper we exploited the random assignment of schools to the teacher evaluation system to estimate the impact of the observation-based evaluation system on student achievement. As such, this paper brings new evidence to bear on the causal effect of a teacher evaluation system based on process-oriented metrics of teacher performance.

8. CONCLUSION

We examine a unique intervention in Chicago Public Schools to uncover the causal impact of an evaluation system, based on structured classroom observations of teacher practice, on school performance. This teacher evaluation pilot included multiple implementation elements—principal training, district support and oversight, and school-level implementation of the principal-teacher classroom observation and conference process—and it is the impact of this system that we report in this paper. Leveraging the random assignment of schools to the EITP pilot intervention, we find there are large short-term effects of classroom observation on school reading performance, and these effects vary by school composition (specifically, school performance and poverty). These effects reflect not only the emphasis of the EITP pilot—to improve a teacher's instructional practice, and linking this structured guidance with student achievement—but also the extent to which CPS supported the implementation of the classroom observation-based evaluation process. Indeed, the nature of implementation fidelity is critically important to consider when assessing the potential impact of any new education initiative.

The implementation of the EITP pilot in Chicago occurred prior to the national attention that is now being placed on the design and use of rigorous teacher evaluation systems. These new teacher evaluation systems, motivated in large part by the federal Race to the Top initiative, incorporate multiple measures of teacher performance, including value-added metrics based on standardized tests and/or teacher-designed assessments, student feedback on teacher performance, classroom observation ratings, peer evaluations, and so on. Compared with the comprehensive evaluation systems being designed and implemented today, the EITP was solely focused on the classroom observation component of teacher evaluation. What is notable about the version of teacher evaluation systems currently evolving in districts throughout the nation, however, is an emphasis on the classroom observation component, with many

systems utilizing the very same observation tool used in CPS under the EITP initiative.

Given that the newest vintage of teacher evaluation systems places significant weight (up to half, in many cases) on the classroom observation component of a teacher's summative rating, this paper provides insight into the potential impact of these metrics on school performance. Indeed, the advantage of this paper is that we are able to isolate the impact that the classroom observation process has on student learning. Going forward, it will be difficult, if not impossible, for researchers to disentangle the impact of classroom observations on student learning from any of the other components of the teacher evaluation system.

While this paper has revealed the causal effect of a classroom observation-based evaluation system on student achievement in a low-stakes environment, a number of important issues remain unexamined. Specifically, what are the mechanisms through which the evaluation pilot produced changes in school performance? For example, did the teacher evaluation pilot realize the intermediate goals of the policy by producing changes in instructional climate or by altering the nature of within-school teacher collaboration, such as the quality of professional conversations about teaching practice? Moreover, do teachers respond to being evaluated, even in a low-stakes context? Specifically, to what extent does a performance evaluation system lead to teacher mobility and turnover? These (and other) questions are currently being pursued by the authors and should shed light on both the black box of teacher evaluation systems aimed at improving teacher instructional practice as well as the potential labor market implications of such systems.

The authors thank Dan Black, Ofer Malamud, Lisa Barrow, Elaine Allensworth, Sue Sporte, Emily Krone, and Sheila Cashman for helpful suggestions on earlier drafts of this paper. The authors benefitted from the suggestions of two anonymous referees and the journal editors, and from editorial assistance from Jennifer Moore. The authors also thank the University of Chicago Consortium on Chicago School Research and staff at Chicago Public Schools for providing access to the data for this research. We would also like to thank the Joyce Foundation and the Spencer Foundation for their continued support of the University of Chicago Consortium on Chicago School Research.

REFERENCES

Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics* 25(1):95–135. doi:10.1086/508733

Allensworth, Elaine, Stephen Ponisciak, and Christopher Mazzeo. 2009. The schools teachers leave: Teacher mobility in Chicago Public Schools. University of Chicago Consortium on Chicago School Research, Research Report.

Bloom, Howard S., Carolyn J. Hill, Alison R. Black, and Mark W. Lipsey. 2008. Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness* 1(4):289–328. doi:10.1080/19345740802400072

Boyd, Donald, Pamela Grossman, Hamilton Lankford, Susanna Loeb, and James Wyckoff. 2009. Who leaves? Teacher attrition and student achievement. CALDER Working Paper No. 23, Urban Institute.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2007. How and why do teacher credentials matter for student achievement? NBER Working Paper No. 12828.

Danielson, Charlotte. 1996. *Enhancing professional practice: A framework for teaching*. Alexandria, VA: Association for Supervision and Curriculum Development.

Dee, Thomas S., and Brian Jacob. 2011. The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and Management* 30(3):418–446. doi:10.1002/pam.20586

Goe, Laura, Courtney Bell, and Olivia Little. 2008. *Approaches to evaluating teacher effectiveness: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality.

Goldhaber, Dan. 2002. The mystery of good teaching. *Education Next* 2(1):50–55.

Goldhaber, Dan. 2007. Everyone's doing it, but what does teacher testing tell us about teacher effectiveness? *Journal of Human Resources* 42(4):765–794.

Hanushek, Eric A., John F. Kain, and Steven G. Rivkin. 2004. Why public schools lose teachers. *Journal of Human Resources* 39(2):326–354. doi:10.2307/3559017

Harris, Douglas N., and Tim R. Sass. 2007. Teacher training, teacher quality, and student achievement. CALDER Working Paper No. 3, Urban Institute.

Ingersoll, Richard M. 2001. Teacher turnover and teacher shortages: An organizational analysis. *American Educational Research Journal* 38(3):499–534. doi:10.3102/00028312038003499

Jackson, Kirabo, and Elias Bruegmann. 2009. Teaching students and teaching each other: The importance of peer learning for teachers. *American Economic Journal: Applied Economics* 1(4):85–108. doi:10.1257/app.1.4.85

Kane, Thomas J., Jonah E. Rockoff, and Douglas O. Staiger. 2008. What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review* 27(6):615–631. doi:10.1016/j.econedurev.2007.05.005

Kane, Thomas J., Eric S. Taylor, John H. Tyler, and Amy L. Wooten. 2011. Identifying effective classroom practices using student achievement data. *Journal of Human Resources* 46(3):587–613. doi:10.1353/jhr.2011.0010

Murnane, Richard, and Jennifer L. Steele. 2007. What is the problem? The challenge of providing effective teachers for all children. *Future of Children* 17(1):15–43. doi:10.1353/foc.2007.0010

National Commission on Excellence in Education (NCEE). 1983. *A nation at risk: The imperative for educational reform*. Washington, DC: U.S. Department of Education.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. Teachers, schools and academic achievement. *Econometrica* 73(2):417–458. doi:10.1111/j.1468-0262.2005.00584.x

Rockoff, Jonah E. 2004. The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review* 94(2):247–252. doi:10.1257/0002828041302244

Rockoff, Jonah E., Brian A. Jacob, Thomas J. Kane, and Douglas O. Staiger. 2011. Can you recognize an effective teacher when you recruit one? *Education Finance and Policy* 6(1):43–74. doi:10.1162/EDFP_a_00022

Sartain, Lauren, Sara R. Stoelinga, Eric Brown, Stuart Luppescu, Kavita K. Matsko, Frances K. Miller, Claire Durwood, Jennie Y. Jiang, and Danielle Glazer. 2011. Rethinking teacher evaluation in Chicago: Lessons learned from classroom observations, principal-teacher conferences, and district implementation. University of Chicago Consortium on Chicago School Research, Research Report.

Sartain, Lauren, Sara R. Stoelinga, and Emily Krone. 2010. *Rethinking teacher evaluation: Findings from the first year of the Excellence in Teaching Project in Chicago Public Schools*. University of Chicago Consortium on Chicago School Research, Policy Brief.

Taylor, Eric S., and John H. Tyler. 2012. The effect of evaluation on teacher performance. *American Economic Review* 102(7):3628–3651. doi:10.1257/aer.102.7.3628

The New Teacher Project (TNTP). 2007. *Teacher hiring, assignment, and transfer in Chicago Public Schools*. Brooklyn, NY: The New Teacher Project.

Watson, J. G., S. B. Kraemer, and C. A. Thorn. 2009. *The other 69 percent*. Washington, DC: Center for Educator Compensation Reform, U.S. Department of Education, Office of Elementary and Secondary Education.

Weisberg, Daniel, Susan Sexton, Jennifer Mulhern, and David Keeling. 2009. *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. Brooklyn, NY: The New Teacher Project.