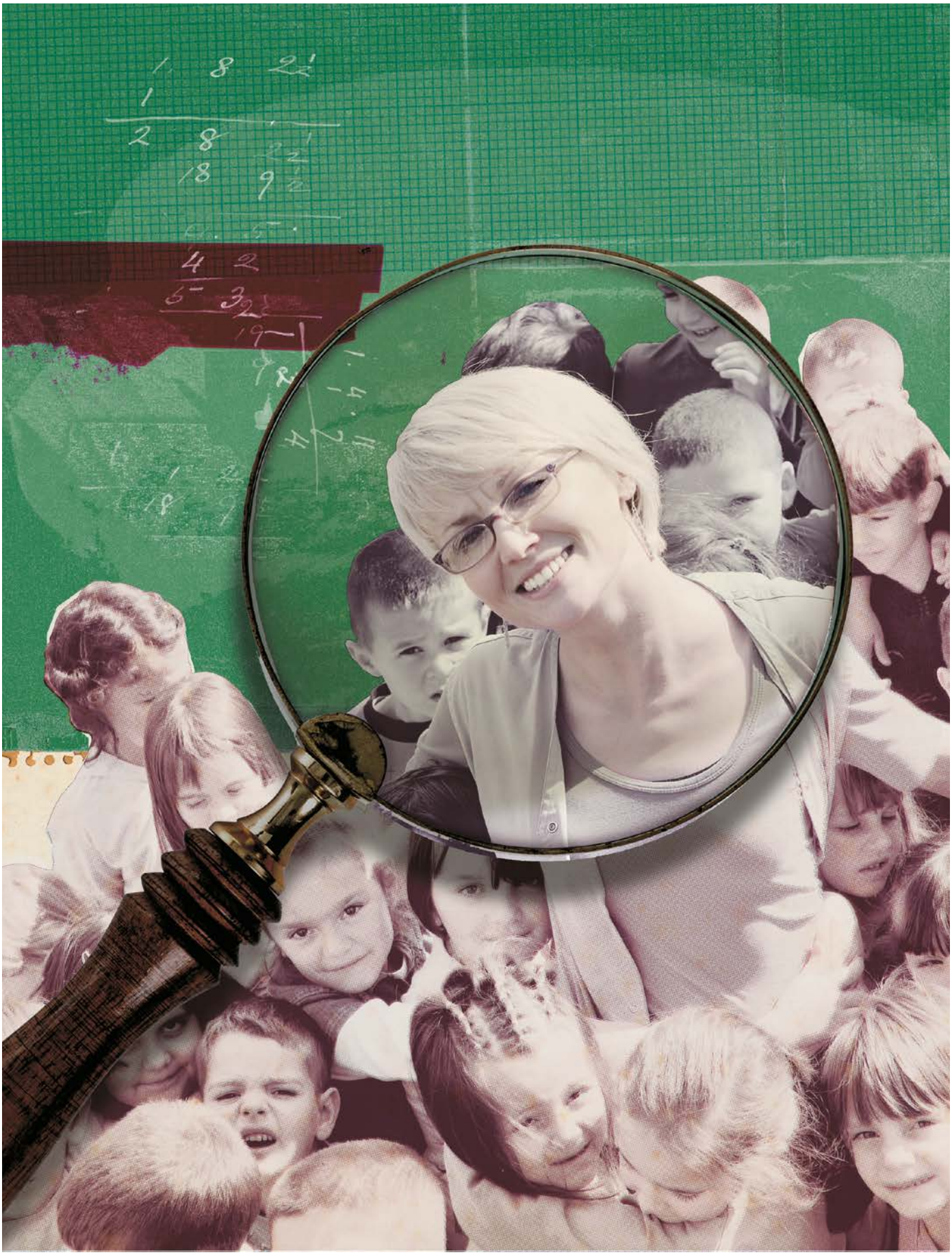


ILLUSTRATION / SARAH HANSON

Does Better Observation Make Better Teachers?

New evidence from a teacher evaluation pilot in Chicago

OF ALL SCHOOL-LEVEL FACTORS related to student learning and achievement, the quality of the student's teacher is the most important. Yet the teacher evaluation systems in use in American school districts historically have been unable to differentiate teachers who improve student learning from lower-performing educators. Many have failed to differentiate teachers at all. A 2009 study by The New Teacher Project found that "satisfactory" or "unsatisfactory" were the only ratings available to school administrators in many districts, and that more than 99 percent of teachers in those districts were deemed satisfactory.

Improving methods for evaluating teacher performance and using the resulting information to change teaching practice has been a focus of recent reform efforts. According to the National Council on Teacher Quality, 32 states and the District of Columbia altered their teacher-evaluation policies in recent years to incorporate multiple methods of assessing and evaluating teachers, spurred in part by the federal Race to the Top competition. And each of the 43 states to which the Obama administration has granted a waiver from No Child Left Behind is now in the process of implementing evaluation systems that employ multiple measures of classroom performance, including student achievement data. These systems differentiate among three or more performance levels and are used to inform personnel decisions.

While much of the debate over these new evaluation systems centers on their use of student test-score data to measure a teacher's "value added" to student learning, classroom observations remain critically important. Most teachers work in grades or subjects in which standardized tests are not administered and therefore will not have a value-added score. Even when students' test scores are available, classroom observations may capture dimensions of teachers' performance that are important but not reflected in those scores. Finally, value-added scores on their own do not tell teachers how they might improve their practice and thereby raise student achievement.

We examine a unique intervention in Chicago Public Schools (CPS) to uncover the causal impact on school performance of an evaluation system based on highly structured classroom observations of teacher practice. An iterative process of observation and conferencing focused on improving lesson planning and preparation, the classroom environment, and instructional techniques should drive positive changes in teacher practice. As teachers refine their skills and learn how best to respond to their students' learning needs, student performance should improve. Recent evidence from Cincinnati Public Schools confirms that providing midcareer teachers with evaluative feedback based on the Danielson Framework for Teaching observation system can promote student-achievement growth in math, both during the school year in which the teacher is evaluated and in the years after evaluation (see "Can Teacher Evaluation Improve Teaching?" *research,* Fall 2012).

The Excellence in Teaching Project (EITP), a teacher evaluation system also based on the Danielson framework, was piloted in Chicago Public Schools beginning in the fall of 2008. Leveraging the random assignment of schools to the EITP intervention, we find large effects of the intervention on school reading performance. The program had the largest impact in low-poverty and high-achieving schools but little or no impact in less-advantaged schools. These effects seem to be a consequence

by MATTHEW P. STEINBERG AND LAUREN SARTAIN

not only of the design and focus of the EITP pilot but also of the extent to which CPS supported the implementation of the new evaluation process. Similar benefits were not observed in schools implementing the same program the following year with less support from the central office, suggesting the importance of sustained support for teacher evaluation reform to translate into improved student performance.

Teacher Evaluation in Chicago Public Schools

For nearly four decades prior to the introduction of the EITP, CPS teachers were observed and evaluated based on a checklist of 19 classroom practices. During a classroom observation of a teacher's lesson, the observer (usually the principal, but sometimes an assistant principal) would check one of three boxes (Strength, Weakness, Does Not Apply) next to each of the practices. The checklist approach was unpopular among both teachers and principals. High-performing teachers believed that 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. The system provided no formal guidance or rubric to either party on what constituted strong or weak performance on any of the checklist practices.

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.

Moreover, there was no direct correspondence between a teacher's ratings on the checklist and the overall evaluation rating, which determined teacher tenure. Overall evaluations also showed little differentiation among teachers. Nearly all teachers (93 percent) received ratings of "Superior" or "Excellent" (the top-two categories in a four-tier rating system). Meanwhile, two-thirds of CPS schools failed to meet state proficiency standards under Illinois's accountability system, and Chicago remained among the nation's lowest-performing urban districts on the National Assessment of Educational Progress.

Dissatisfaction with the evaluation system led CPS leadership under then CEO Arne Duncan to develop the EITP in partnership with the Chicago Teachers Union (CTU), beginning in 2006. A joint CPS–CTU committee met together over two years to negotiate the details of the evaluation pilot. In the summer of 2008, just prior to implementation, the district and union disagreed on whether the ratings teachers received under the EITP would be used for teacher accountability purposes, such as tenure decisions. The district nonetheless moved forward with the pilot to implement formative, ongoing assessments for teachers that would provide them with structured feedback on their instructional practices.

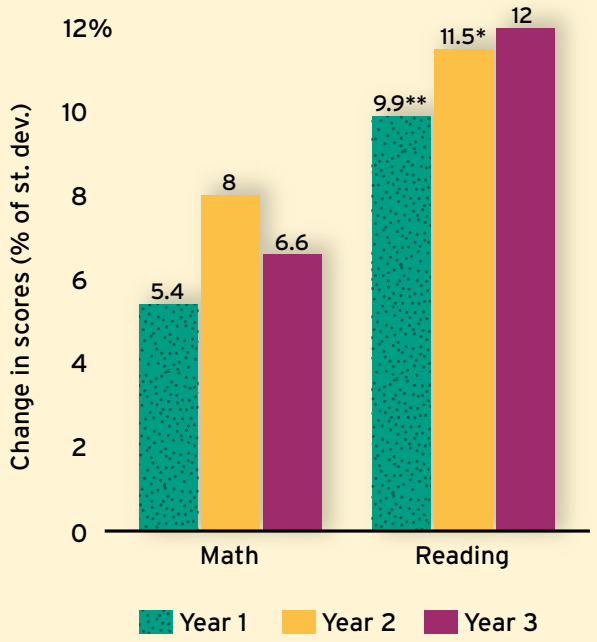
The classroom observation process had occurred formally (if superficially) twice a year for all teachers, irrespective of tenure status, as part of the district–union teacher contract. While maintaining this schedule, the EITP changed the process significantly. First, principals and teachers engaged in a brief (15- to 20-minute) pre-observation conference during which they reviewed the rubric. The conference also gave the teacher an opportunity to share any information about the classroom with the principal, such as issues with individual students or specific areas of practice about which the teacher wanted feedback. During the 30- to 60-minute lesson that followed, the principal was to take detailed notes about what the teacher and students were doing. After the observation, the principal was expected to match classroom observation notes to the Danielson framework rubric in order to rate teacher performance in 10 areas of instructional practice.

The Danielson framework delineates four levels of performance (Unsatisfactory, Basic, Proficient, and Distinguished) across four domains, of which the EITP focused on two: Classroom Environment and Instruction. Within a week of the observation, the principal and teacher conducted a postobservation 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.

The 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. In 2008–09, the first year of implementation, 44 participating principals received approximately 50 hours of training and support, with three days of initial training during the summer and follow-up sessions throughout the school year. The initial training covered the use of the Danielson framework to rate teaching practice, methods for collecting evidence, and best practices for conducting classroom observations. The follow-up sessions consisted of seven monthly meetings in which principals brought materials from classroom observations that they had conducted and engaged in small-group discussion with their colleagues. Four additional half-day trainings during the school year provided an opportunity for principals to update their

Reading Boost (Figure 1)

Schools that implemented the Excellence in Teaching Project in the first year saw large improvements in reading achievement relative to schools that launched the program in the second year.



* Effect is statistically significant at the 90 percent confidence level
** Effect is statistically significant at the 95 percent confidence level

SOURCE: Authors' calculations

understanding and use of the rubric for evaluating teachers.

Principals also received additional one-on-one 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. 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 principals in making adjustments to their own ratings of teacher performance.

Forty-four schools participated in EITP in the first year. These 44 Cohort 1 schools continued to take part in the second year, and an additional 48 schools (Cohort 2) implemented EITP for the first time. The extent of principal training and support for the 48 new schools differed dramatically from Cohort 1, however. In their first year, Cohort 2 principals received just two days of initial training on how to collect evidence on 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 had in their first year of implementation. Although Cohort 2 principals could request technical assistance from EITP central-office staff, these principals did not have access to the ongoing technical support and oversight that Cohort 1 principals received. 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.

Data

Data for this study consist of CPS administrative, personnel, and test-score information from the 2005–06 school year to the 2010–11 school year. As the intervention occurred at the school level, we used school-level averages of all student-level and teacher-level data records. 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. We also use school-level characteristics such as student enrollment levels and the distribution of race/ethnicity, gender, students qualifying for free or reduced-price lunch, and special education students, which were generated from student-level CPS data files. 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.

Forty-four schools participated in EITP in the first year, and in the second year, an additional 48 schools implemented EITP.

Our primary outcome variable is student achievement as measured by performance on standardized tests. Students in

Illinois take the Illinois Standards Achievement Test (ISAT) in reading and mathematics in grades 3 through 8, usually in March of each school year. We use a school-level measure that has been standardized across the sample of schools included in our analysis, taking into account the various grade configurations in different schools.

Methodology

We take advantage of a unique randomized control trial design. CPS, in partnership with CCSR, selected four elementary-school instructional areas (of the 17 elementary zones in the city at the time) that would implement the EITP. 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 first year of EITP (Cohort 1). Schools with first-year principals and those slated for closure in the spring of 2009 were excluded from the sample prior to randomization.

Schools that were not selected to participate in the first year implemented the program the following school year (Cohort 2). The randomization process resulted in 44 Cohort 1 schools and 49 Cohort 2 schools (the latter number fell to 48 due to the unexpected closure of one school). Our data indicate that the randomization procedure worked as desired. On average, the Cohort 1 and Cohort 2 schools were very similar in terms of both student and teacher characteristics as well as school working conditions.

We measure the initial impact of the EITP on a school's math and reading achievement by comparing student achievement between the Cohort 1 and Cohort 2 schools at the end of the 2008–09 school year, during which Cohort 1 schools implemented the EITP but Cohort 2 schools did not. To increase the precision of our results, we control for student enrollment, the proportion of female students, the proportion of students by race/ethnicity, the proportion of special education students, the proportion of students receiving free or reduced-price lunch, and average prior achievement.

In its first year, the EITP increased student achievement in reading in the Cohort I schools by an effect equivalent to closing one-quarter to one-half of the performance gap between weak schools and average schools in large urban districts like Chicago.

Results

In its first year, the EITP increased student achievement in the Cohort 1 schools by 5.4 percent of a standard deviation in math and 9.9 percent of a standard deviation in reading, relative to the Cohort 2 schools. The effect on reading scores is statistically significant, but the effect on math scores is not. The reading effect is significant not just statistically but also in size. A 10 percent of a standard deviation effect size is equivalent to closing 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 like Chicago.

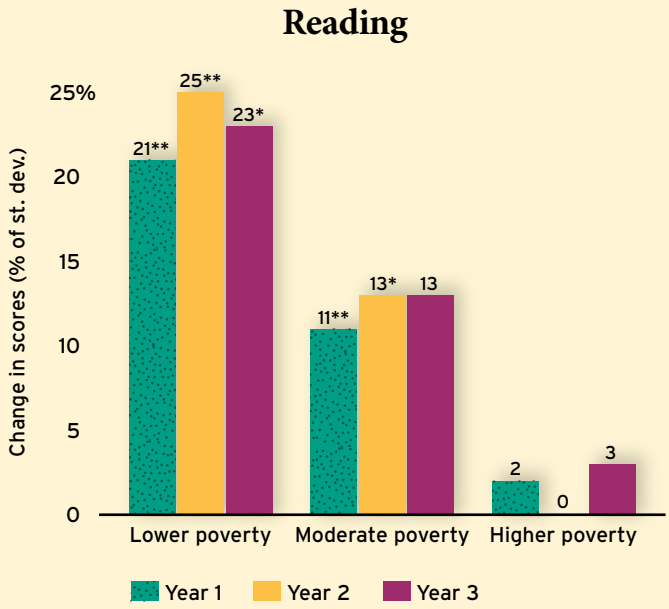
In the second year, as Cohort 2 schools implemented EITP, we might have expected the difference between the two groups of schools to shrink or even disappear as the Cohort 2 schools benefited from the same program that had a positive impact on Cohort 1 schools the prior year. We find, however, that the difference in student achievement between the two groups of schools persisted over time. Figure 1 shows that the math effect of 5.4 percent increased to 8 percent in year two and was 6.6 percent in year three. For reading, the first-year effect of 9.9 percent grew to 11.5 and 12 percent in the second and third years.

More-advantaged schools—those with fewer students eligible for free or reduced-price lunch and those with higher initial student achievement—benefited the most from the program. On average across all schools, 83 percent of students received free or reduced-price lunch. The effect of EITP at lower-poverty schools—those with just 60 percent of students receiving free or reduced-price lunch—was double the effect for the full sample, at more than 20 percent of a standard deviation (see Figure 2). On the other end of the distribution, there was no detectable EITP effect at higher-poverty schools. This differential effect persisted into the second and third years of the intervention, after Cohort 2 schools implemented the program.

We find similar differential effects on math by school poverty level, with a statistically significant positive effect for lower-poverty schools, even though the average effect across all schools was not distinguishable from zero. We also find evidence that schools with higher student achievement before the start of the EITP benefited the most from the program. We do not, however, find any consistent evidence that the effect of the program was related to the racial composition or share of special education students in the school.

Impact Variations (Figure 2)

Among the schools implementing the Excellence in Teaching Project in the first year, lower-poverty schools saw the greatest impact on reading scores.



* Effect is statistically significant at the 90 percent confidence level
** Effect is statistically significant at the 95 percent confidence level

NOTE: In lower-poverty schools, 60 percent of students are eligible for free or reduced-price lunch, in moderate-poverty schools 83 percent of students are eligible, and in higher-poverty schools virtually all students are eligible.

SOURCE: Authors' calculations

Explaining Differential Impacts

Why did the EITP only improve achievement in certain schools and only in the first year? The EITP represented a dramatic departure from the existing teacher-evaluation system in Chicago and relied on the human capital that already existed in the schools to generate improvements in school performance. Its efficacy depended on principals' capacity to provide targeted instructional guidance, 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 the new system.

The pilot forced principals to make significant changes to how they conducted classroom observations and conferences with teachers. The intervention itself was time-intensive for the principals, who were required to participate in extensive training pre-intervention. Principals also had to rate teachers on the new evaluation framework, and work with them in pre- and postobservation conferences to develop strategies to improve their instructional practice. On average, CPS principals reported that they spend about six hours per teacher during each formal observation cycle.

The principals' role evolved from pure evaluation to a dual role in which, by incorporating instructional coaching, the principal served as both evaluator and formative assessor of a teacher's instructional practice. It seems reasonable to expect that more-able principals could make this transition more effectively than less-able principals. A very similar argument can be made for the demands that the new evaluation process placed on teachers. More-capable teachers are likely more able to incorporate principal feedback and assessment into their instructional practice.

Our results indicate that while the pilot evaluation system led to large short-term, positive effects on school reading performance, these effects were concentrated in schools that, on average, served higher-achieving and less-disadvantaged students. For high-poverty schools, the effect of the pilot is basically zero.

We suspect that this finding is the result of the unequal allocation of principals and teachers across schools as well as additional demands placed on teachers and principals in more disadvantaged schools, which may impede their abilities to implement these types of reforms. For example, if higher-quality principals and teachers are concentrated in higher-achieving, lower-poverty schools, it should not be surprising that a program that relies on high-quality principals and teachers has larger effects in these schools. In addition, less-advantaged schools with, on average, harder-to-serve student populations, may require additional supports for these kinds of interventions to generate improvements in student learning similar to those of more-advantaged schools.

Varied Implementation

School-level implementation is critically important for the success of any new educational intervention. As discussed above, the extent of principal training and district-level support varied dramatically for Cohort 1 and Cohort 2 schools. We speculate that district support also played an important role in explaining the large positive effect for Cohort 1 and the null effect for Cohort 2.

Leadership turnover in CPS led to a decline in institutional and district support for EITP between the first and

second years of the pilot program. When the pilot 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 program 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 to serve as U.S. secretary of education. While Duncan's arrival in Washington in early 2009 was followed by a national emphasis on refining teacher evaluation systems, his departure from Chicago marked a move away from the rigorous year one implementation of the EITP pilot. The incoming administration deemphasized the teacher evaluation pilot and instead focused on performance monitoring, data usage, and accountability.

When the EITP expanded to include the Cohort 2 schools in 2009, doubling the number of schools implementing the pilot, the budget for district support of the program did not increase. This limited the amount of support the central office could provide to principals, which we suspect reduced the fidelity with which the pilot was implemented and in turn weakened the intervention. CPS central-office staff responsible for EITP oversight and school-level implementation indicated that there was a significant decrease in both CPS staff and budgetary resources dedicated to Cohort 2 principals in comparison to 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 system implementation. 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 it. As a result, there is little evidence that the Danielson framework was used in any systematic way in year three. Our results are consistent with strong implementation in year one and weak or no implementation in subsequent years.

Chicago's decision to abandon the EITP pilot, after supporting it fully for just one year, illustrates the difficulty urban school districts have in sustaining large-scale policy changes.

Conclusion

The implementation of the EITP pilot in Chicago occurred prior to the nationwide shift toward more rigorous teacher-evaluation systems. These new teacher-evaluation systems incorporate multiple measures of teacher performance, including value-added metrics based on standardized tests or teacher-designed assessments and, in some cases, student feedback on teacher performance and peer evaluations. Unlike these systems, the EITP was focused solely on classroom observation. What is notable about the version of teacher evaluation systems currently evolving in districts throughout the nation, however, is the continued emphasis on classroom observations, with many systems employing the same observation tool used in CPS under the EITP initiative.

A number of important issues remain unexamined. Specifically, what are the mechanisms through which the evaluation pilot produced improvements in school performance? For example, did the teacher evaluation pilot produce changes in instructional climate or alter the nature of within-school teacher collaboration? To what extent does a performance evaluation system alter teacher mobility and turnover patterns? Answers to these and other questions will shed light on how teacher evaluation systems might improve instructional practice as well as their implications for the teacher labor market.

Chicago's decision to abandon the EITP pilot, after supporting it fully for just one year, illustrates the difficulty urban school districts have in sustaining large-scale policy changes that require ongoing support from the central office and significant investment on the part of educators in specific schools. In this case, the program had considerable promise. In the fall of 2012, CPS launched a new teacher-evaluation program in order to comply with the Illinois Performance Evaluation Reform Act, which requires that indicators of student growth be a "significant factor" in teacher evaluation. Called REACH (Recognizing Educators Advancing Chicago Students), the new program also uses the Danielson framework for the classroom observation component.

Matthew Steinberg is assistant professor of education at the University of Pennsylvania Graduate School of Education. Lauren Sartain is research analyst at the Consortium on Chicago School Research at the University of Chicago. This article is based on a forthcoming study in Education Finance and Policy.