A LASTING IMPACT

High-stakes teacher evaluations drive student success in Washington, D.C

TEACHERS MATTER—and some matter more than others. That recognition has driven a tidal wave of controversial policy reforms over the past decade, rooted in new evaluation systems that link teachers' ratings and, in some cases, their pay and advancement to evidence of classroom practice and student learning. Two out of three U.S. states overhauled teacher evaluations between 2009 and 2015, supported by federal incentives such as Race to the Top and Teacher Incentive Fund grants, as well as No Child Left Behind Act waivers.

What is the impact, so far, of these reforms? One common narrative would indicate a flop. Most states adopting new evaluation systems saw little change in the share of teachers deemed less than effective, arguably limiting their potential to address underperformance. Meanwhile, a broader backlash against reform, fueled by concerns about overreliance on standardized tests, the accuracy of new evaluations, and the efficacy of performance-based incentives, has led some states to reverse course. Congress ultimately chose to exclude any requirements about teacher evaluation policies from the Every Student Succeeds Act of 2015, dashing some reformers' hopes for a federal mandate.

A closer look at one high-stakes evaluation system, however, shows the positive consequences such systems can have for students. Since 2012, we have been studying IMPACT, a seminal effort by the District of Columbia Public Schools (DCPS) to link teacher retention and pay to their performance. Under IMPACT, the district sets detailed standards for high-quality instruction, conducts multiple observations, assesses individual performance based on evidence of student progress, and retains and rewards teachers based on annual ratings. Looking across our analyses, we see that under IMPACT, DCPS has dramatically improved the quality of teaching in its schools—likely contributing to its status as the fastest-improving large urban school system in the United States as measured by the National Assessment of Educational Progress.

Such reforms are often considered politically impossible, and the effort in DCPS shows the potential fallout. The district's controversial chancellor, Michelle Rhee, resigned a year after IMPACT launched, when the mayor who appointed her, Adrian Fenty, lost a reelection

by THOMAS DEE and JAMES WYCKOFF

Kaya Henderson was Chancellor of the District of Columbia Public Schools from 2010–2016.

PHOTOGRAPH / SARAH L. VOISIN, THE *WASHINGTON POST*

bid in a campaign focused on school reform.

But IMPACT outlasted them both, to the benefit of students. DCPS dismissed the majority of very low performing teachers and replaced them with teachers whose students did better, especially in math. Other low-performing teachers were 50 percent more likely to leave their jobs voluntarily, and those who opted to stay improved significantly, on average, the following year. High-performing teachers improved their performance as well, especially those within reach of the significant financial incentive created by the system. Certainly, improvement was not universal, and some very good teachers decided to leave the district. Nonetheless, our analysis finds that improved teaching was common and that student achievement increased as a result.

The DCPS story shows that it may be politically challenging to adopt high-stakes evaluation systems, but it is not impossible. And it shows that well-designed and carefully implemented teacher evaluations can serve as an important district improvement strategy—so long as states and districts are also willing to make tough, performance-based decisions about teacher retention, development, and pay.

A More-Rigorous Approach

Teacher evaluations came into focus in 2009 alongside a growing research consensus that teacher quality has dramatic, long-lasting impacts on student success. DCPS was in the vanguard of these efforts, and was one of the first school districts in the country to link individual performance evaluations to high-stakes decisions about pay and retention when it launched IMPACT in the 2009–10 school year.

IMPACT articulates clear standards for effective instruction, provides instructional coaches to help teachers meet those standards, and evaluates teacher performance based in large part on direct, structured observations of teachers' classroom practices in addition to evidence of student learning. IMPACT's features are broadly consistent with emerging best-practice design principles informed by the Measures of Effective Teaching project, and are intended to drive improvements in teacher quality and student achievement (see "Capturing the Dimensions of Effective Teaching," *features,* Fall 2012).

All teachers receive a single IMPACT score ranging from 100 to 400 points at the end of each school year based on classroom observations, measures of student learning, and commitment to the school community.

Under IMPACT, all teachers receive a single score ranging from 100 to 400 points at the end of each school year based on classroom observations, measures of student learning, and commitment to the school community. However, the components of the scores and the weights assigned to them vary, depending on what measures of student performance are available, and DCPS has adjusted their composition over time. Below, we describe the components of IMPACT as it existed in its first three years.

All teachers were evaluated by five structured classroom observations aligned to the district's Teaching and Learning Framework, which defined domains of effective instruction, such as leading well-organized, objective-driven lessons; checking for student understanding; explaining content clearly; and maximizing instructional time. Of the five observations, three were conducted by an administrator and two by "master educators" who traveled between several schools; only the first observation was announced in advance.

The weight of these observation scores varied, depending on whether a teacher also had a value-added score based on her students' performance on standardized tests. For those teachers—who led reading or math classrooms in grades 4–8 and accounted for less than one in five DCPS teachers—observations were worth 35 percent and value-added was worth 50 percent. For the majority of DCPS teachers who did not have value-added scores, observations counted for 75 percent of the total IMPACT score. An additional 10 percent was based on progress toward goals for student learning, which teachers and administrators set each year.

A teacher's contribution to a school's community, as assessed by the principal, was worth 10 percent of the overall evaluation score, while the final 5 percent was based on a measure of the value-added to student achievement for the school as a whole. Teachers could also have points deducted for failing to meet expectations

for attendance, punctuality, and adherence to other policies and procedures.

Telling Teachers Apart

Unlike typical teacher-evaluation systems, IMPACT creates substantial differentiation in ratings. In 2009–10 and 2010–11, 14 percent of teachers were rated "highly effective," 69 percent of teachers were rated "effective," 14 percent were judged "minimally effective," and another 2 percent were deemed "ineffective."

The system also sets swift consequences for teachers with very low or very high scores. Teachers rated "ineffective" are dismissed with rare exception, a rule that caused more than 200 teachers to exit DCPS in IMPACT's first year. Those rated "minimally effective" are given one year to raise their score above the "effective" threshold; if they do not, they are dismissed.

On the other side of the spectrum, teachers are eligible for a bonus of up to $25,000 in each year they are rated "highly effective;" those with "highly effective" scores two or more years in a row can earn increases in their base pay of up to $27,000. Teachers working in high-poverty schools teaching high-need subjects are eligible for the largest pay increases.

These design features create sharp incentive contrasts for teachers with scores on either side of the "effective" threshold, because scoring above it removes the threat of dismissal. And they create an incentive for teachers with scores just below the "highly effective" threshold, because scoring above it makes them eligible for a significant increase in pay.

Our research on IMPACT sought to understand whether those incentives impact teachers' performance and retention—and whether those impacts, if any, improve student outcomes. We based our analysis on administrative data on all DCPS teachers and their students over the first three years of IMPACT: 2009–10, 2010–11, and 2011–12. We limited our sample to general-education teachers working in schools serving K–12 students. We also drew on an additional year of data, from the 2012–13 school year, in assessing IMPACT's effects on student achievement in tested grades and subjects.

Unlike typical teacher-evaluation systems, IMPACT creates substantial differentiation in ratings.

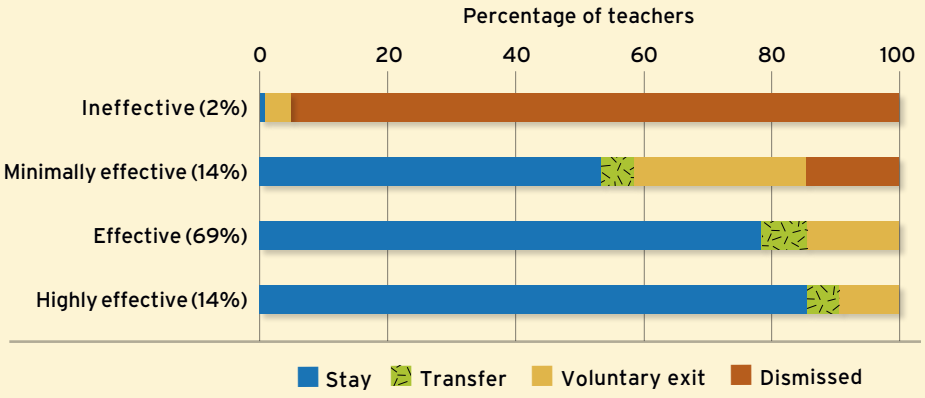
Teacher Retention and Performance

DCPS teachers were retained differently depending on their IMPACT ratings. By the program's third year, 95 percent of teachers deemed "ineffective" in the system's first two years had been dismissed (see Figure 1). Overall, 3.8 percent of all teachers in the district were let go as a result of being rated "ineffective" once or after earning two consecutive "minimally effective" ratings under IMPACT between 2009–10 and 2011–12.

Under IMPACT, lower-rated teachers were

Differential Teacher Retention Under IMPACT (Figure 1)

Ninety-five percent of teachers identified as "ineffective" under IMPACT were dismissed. Among those identified as "minimally effective," 14 percent were dismissed and 27 percent voluntarily left DCPS. Among "effective" and "highly effective" teachers, the majority stayed at DCPS.



NOTE: Percentages accompanying rating categories represent the percentage of DCPS teachers in each category. IMPACT ratings are based on performance during the 2009-10 and 2010-11 school years, with retention outcomes observed in the 2010-11 and 2011-12 school years. Units of observation are teacher years, so teachers may be observed more than once. An "other" retention category, which is always less than 2 percent of any IMPACT rating group, is omitted.

SOURCE: Authors' calculations

also far more likely to exit DCPS voluntarily. Among all teachers rated "minimally effective," 27 percent voluntarily left the district, compared to 14 percent of teachers rated "effective" and 9 percent of teachers rated "highly effective." "Minimally effective" teachers whose scores were closest to the "effective" threshold were less likely to leave than those with lower scores; about one in four teachers whose scores were within 25 points of the "effective" threshold chose to leave their jobs, compared to about one in three whose scores were more than 25 points below. In other words, teachers under threat of dismissal were more likely to voluntarily leave than teachers not subject to this threat, and those who scored furthest from the "effective" threshold were even more likely to go.

These patterns are consistent with a restructuring of the teacher workforce in response to the incentives embedded in IMPACT, but other explanations are possible. For example, some studies have found that less-effective early-career teachers are more likely to exit than more-effective novice teachers, even in the absence of high-stakes evaluations. In addition, IMPACT scores for teachers in their first two years of teaching average 17 points less than those with three or more years of experience. Such considerations raise doubts about how to interpret the relationship between IMPACT and retention. Are we observing the effects of incentives or merely behavior that would have occurred in the absence of IMPACT?

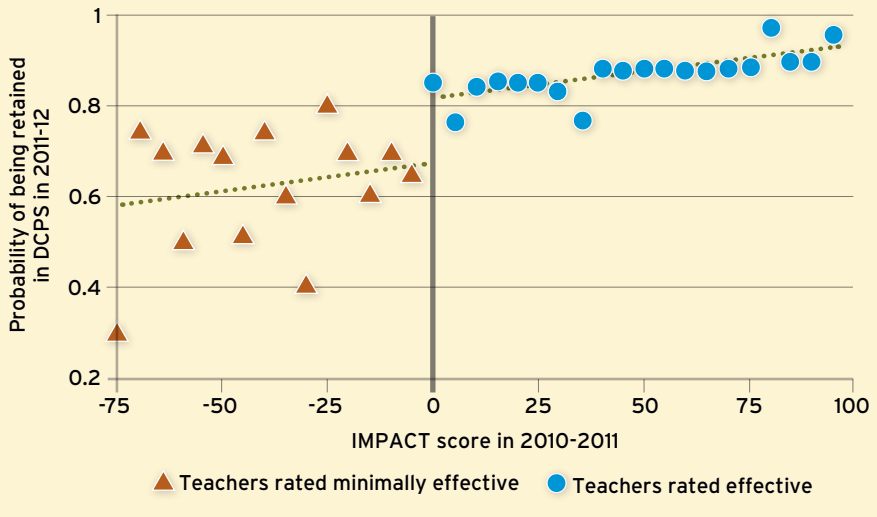
The system sets swift consequences for teachers with very low or very high scores. Teachers rated "ineffective" are dismissed with rare exception, and on the other side of the spectrum, teachers rated "highly effective" are eligible for financial incentives.

To isolate the causal impact of performance-based incentives on teacher quality, we compare outcomes among teachers whose initial IMPACT scores placed them near the thresholds between categories: "minimally effective/effective" and "effective/highly effective." We assume that whether a teacher scores just above or just below a certain threshold is essentially random, which is supported both by our knowledge of how DCPS calculates IMPACT scores and additional analyses confirming that teachers on either side of each threshold are quite similar in terms of experience and other characteristics. We look at teacher retention during the second year of IMPACT, when the threat of dismissal for "minimally effective" teachers became newly credible and financial incentives for "highly effective" teachers could be made permanent. We also examine performance among returning teachers in the system's third year. By comparing teacher attrition and performance on each side of the performance cutoffs, we can get a better sense of how the threat of dismissal or prospect of a raise affects teachers' behavior.

We first use this method to measure the effects on teachers scoring directly above or below the "minimally effective/effective" threshold in 2010–11. We find that teachers near the threshold who received their first "minimally effective" rating at this time were considerably more likely to exit DCPS voluntarily, with retention dropping by 10 percentage points in 2011–12 (see Figure 2). In other words, IMPACT's minimally effective rating increased the attrition of lower-performing teachers from 20 percent

Teachers Rated "Minimally Effective" More Likely to Leave DCPS (Figure 2)

At the threshold where IMPACT scores result in a rating of "minimally effective," teacher retention drops by more than 10 percentage points.



SOURCE: Authors' calculations

to 30 percent, an increment of 50 percent.

Meanwhile, the teachers in this "minimally effective" group that did return to DCPS the following school year improved their performance by roughly 11 points on the IMPACT scale. This sharp increase in teacher performance is about 0.27 of a standard deviation (see Figure 3) and suggests that previously low-performing teachers who opted to remain on the job undertook successful efforts to improve.

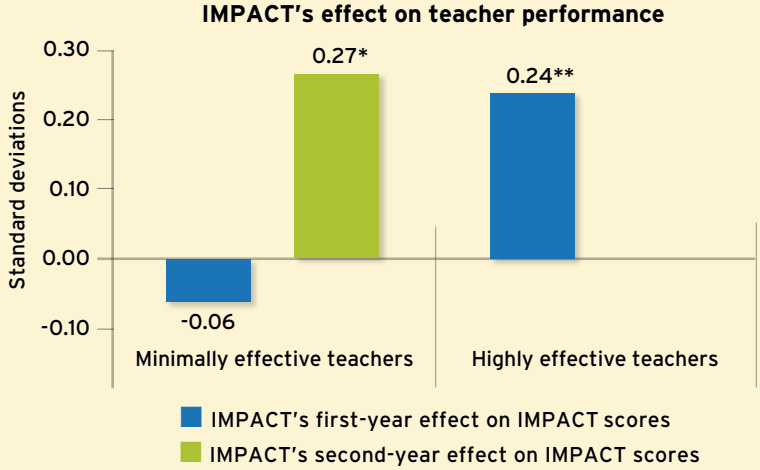
Notably, the effects of a minimally effective rating on retention and performance occurred at the end of IMPACT's second year, when the political credibility of the reform had been affirmed by the appointment of Kaya Henderson as chancellor and by the first instance in which teachers (roughly 140) were fired for having two consecutive "minimally effective" ratings. In contrast, following its first year of implementation, IMPACT had no statistically detectable effect on teacher retention or performance. Given the contentiousness of these policies and the political volatility associated with Mayor Fenty's loss and Chancellor Rhee's resignation, teachers may have reasonably doubted the staying power of IMPACT's performance-based dismissal threats, opting not to respond to its incentives. We are just speculating regarding the cause, but the important lesson is that policies intended to induce significant behavioral responses may need time to take hold.

In contrast, the financial incentives for consistently high-performing teachers (i.e., jumping ahead on the salary schedule) appear to have been immediately credible. Among highly rated teachers who scored very close to the eligibility threshold for a permanent pay increase, retention increased by roughly 3 percentage points, though this effect was not statistically significant. Their performance in 2011–12 improved by a statistically significant 10.9 points, or 0.24 of a standard deviation.

These performance gains imply increases of approximately 5 percentile points in the distribution of teacher performance among lower-rated teachers, and 7 percentile points among highly rated teachers. Comparing these figures to the typical improvement of a novice teacher during her first three years in the classroom, we find that the gain of 12.6 points for lower-rated teachers is 52 percent of this three-year gain, and the gain of 10.9 points for highly rated teachers is 41 percent of the three-year gain.

Teacher Performance Boost (Figure 3)

A "minimally effective" rating in IMPACT's first year had small and statistically insignificant effects on subsequent teacher performance (relative to an "effective" rating). However, in IMPACT's second year, teachers rated "minimally effective" who chose to remain in DCPS improved their performance by 0.27 standard deviations compared teachers rated "effective." Meanwhile, observed in IMPACT's first year, the financial incentives available to teachers on the "highly effective" side of the threshold improved subsequent teacher performance by 0.24 standard deviations.



* Statistically significant at the 95% confidence level

** Statistically significant at the 99% confidence level

NOTE: First-year effects are the 2009-10 IMPACT rating effects on outcomes observed in 2010-11, and second-year effects are the 2010-11 IMPACT rating effects on outcomes observed in 2011-12. This analysis includes only the first-year effects to assess the effect of receiving a rating of "highly effective" for the first time, which triggered a bonus of up to $25,000. Including teachers rated "highly effective" in the second year could also include teachers receiving a permanent salary boost triggered by repeated high ratings. However, including data from IMPACT's second year leave our results qualitatively unchanged.

SOURCE: Authors' calculations

Teacher Turnover and Student Achievement

IMPACT's effects also depend on the direct impact of teacher turnover and the quality of newly hired teachers. Teacher turnover is often assumed to have a universally negative influence on school quality, and replacing teachers in schools with high rates of turnover can place strong demands on district recruitment efforts. Not all turnover is the same, however. Under IMPACT, a substantial fraction of teacher turnover consists of lower-performing teachers who were purposefully compelled or encouraged to leave, which potentially alters the distribution of teacher effectiveness among exiting teachers.

To determine the effect of teacher turnover on student achievement under IMPACT, we examine the year-to-year changes in school-grade combinations with and without teacher turnover. In other words, what was the change in test scores for 4th graders from year to year at a school that had teacher turnover in that grade compared to the change in test scores between 4th graders at a school that did not have teacher turnover in that grade?

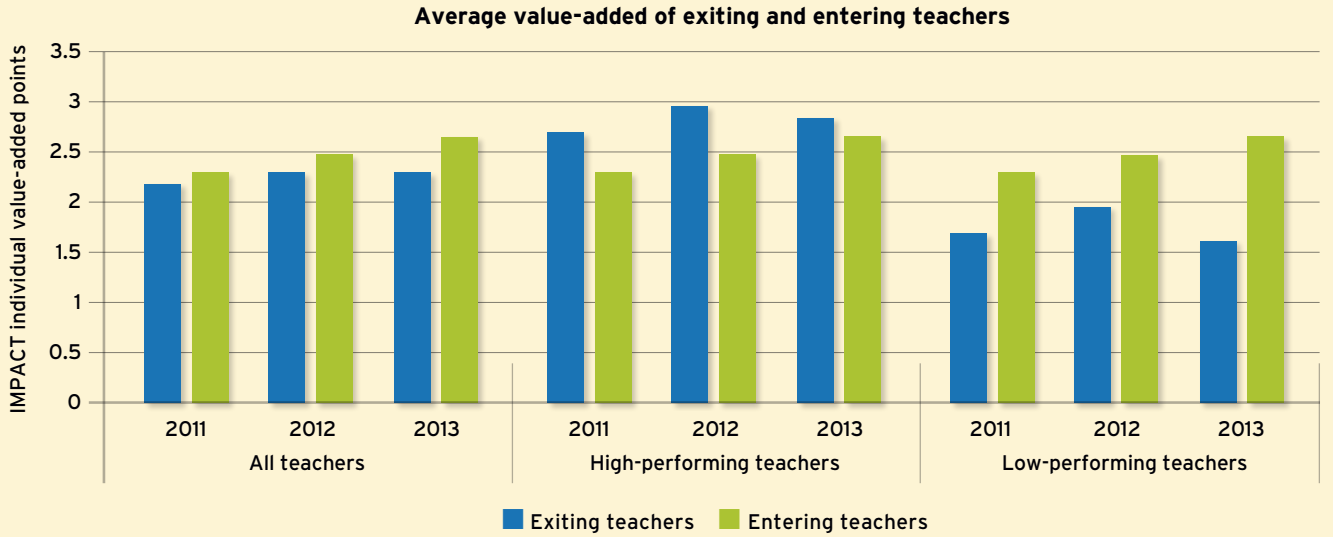
In math, the exit of low-performing teachers is estimated to improve student achievement by 0.21 of a standard deviation—between one-third and two-thirds of a year of learning, depending on grade level.

We find that the overall effect of teacher turnover in DCPS at worst had no adverse effect on student achievement and, under reasonable assumptions, improved it. This average combines the negative, but statistically insignificant, effects of exits of high-performing teachers with the very large improvements in student achievement resulting from the departures of low-performing teachers. Figure 4 illustrates these patterns by comparing the average value-added scores of new DCPS teachers entering the district with those who had exited the previous year.

Looking directly at impacts on student achievement, and carefully controlling for a variety of potentially confounding factors,

Tracking Teacher Quality (Figure 4)

Teacher turnover under IMPACT improved teacher quality. Although high-performing teachers who left DCPS were replaced by somewhat less-effective teachers, on average, when low-performing teachers left DCPS, they were replaced by much more effective teachers, on average.



NOTE: Results for 2011, for example, indicate the average score for all teachers who exited at the end of 2009–2010 compared with all those entering in 2010–2011. Exits include teachers who retired, resigned, or were terminated. Teachers leaving schools that closed are excluded. Average teacher value-added was converted to a 0-4 scale using a conversion table. Data are for teachers who are matched to students with math achievement scores.

SOURCE: Authors' calculations

we find results quite consistent with these averages (see Figure 5). For example, turnover among low-performing teachers leads to improved student achievement in both math and reading. In math, the exit of low-performing teachers is estimated to improve student achievement by 0.21 of a standard deviation—between one-third and two-thirds of a year of learning, depending on grade level. In reading, student achievement is estimated to increase by 0.14 of a standard deviation.

We also compare differences in turnover and their impact on student achievement at high- and low-poverty schools. Overall, we find that high-poverty schools appear to improve as a result of teacher turnover, though as in all schools, not all turnover is the same.

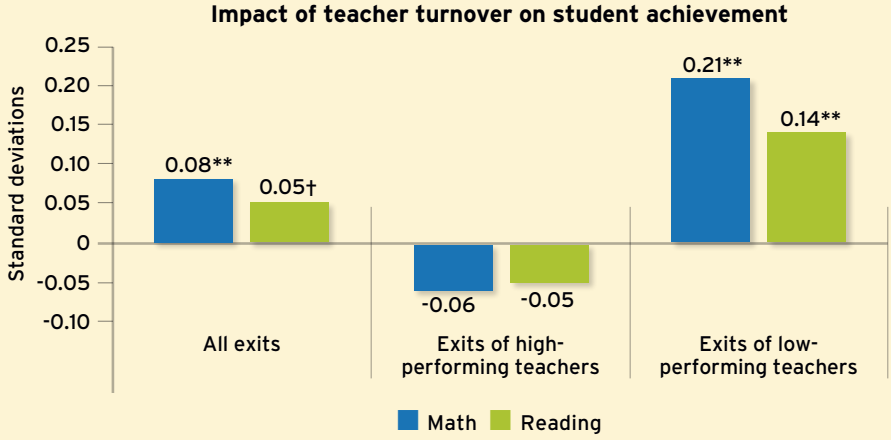
In high-poverty schools, we estimate that the overall effect of all teacher turnover on student achievement is 0.08 of a standard deviation in math and 0.05 of a standard deviation in reading. In comparison, in low-poverty schools, the estimated effects of turnover are close to zero.

However, 40 percent of teacher turnover in high-poverty schools is among low-performing teachers—and our estimates indicate consistently large gains for students when these teachers exit these schools. In math, student achievement improves by 0.21 of a standard deviation, with teacher quality improving by 1.3 standard deviations. In reading, student achievement improves by 0.14 of a standard deviation, with an improvement of 1.0 standard deviations in teacher quality.

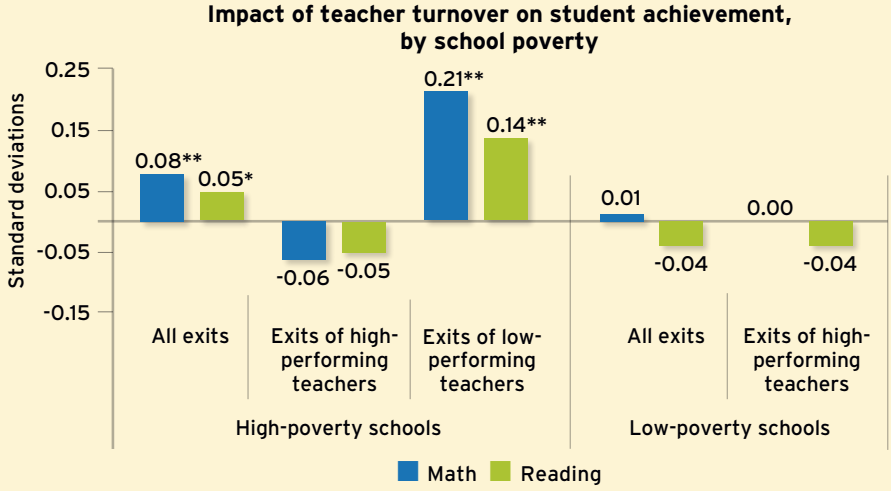
In sum, DCPS has been able to replace low-performing teachers in high-poverty schools with teachers who are substantially more effective. DCPS also appears to be quite capable of replacing exiting high-performing teachers in low-poverty schools with comparable teachers.

Achievement Gains from Turnover (Figure 5)

(5a) The exit of DCPS teachers under IMPACT improved student achievement in both math and reading. The turnover of high-performing teachers had a small, negative, but statistically insignificant effect on achievement in both subjects, while the departure of low-performing teachers substantially increased student achievement, especially in math but also in reading.



(5b) The improvement in student achievement from teacher turnover under IMPACT was primarily concentrated at high-poverty schools.



† Statistically significant at the 90% confidence level
* Statistically significant at the 95% confidence level
** Statistically significant at the 99% confidence level

NOTE: "High-poverty" schools are those where the percentage of students eligible for free and reduced price lunch is at least 60 percent. "High-performing" teachers are those rated effective or highly effective under IMPACT; "low-performing" teachers are those rated ineffective or minimally effective under IMPACT. Data are insufficient to generate estimates for the effect of low-performer exits in low-poverty schools.

SOURCE: Authors' calculations

With the exception of math achievement in one year (2011–12), however, the effect of this turnover on student achievement is not statistically significant.

Implications

The high stakes associated with IMPACT have been a source of contention, both within the District of Columbia as well as in broader discussions of education policy. While the importance of highly effective teachers is uncontroversial, the manner in which schools and districts identify, reward, and retain these educators is.

Our studies provide clear evidence that IMPACT's exceptionally high-powered, individually targeted incentives linked to performance influence retention and performance in desirable ways. This multiple-measures system boosts performance among teachers most immediately facing consequences for their ratings, and promotes higher rates of turnover among the lowest-performing teachers, with positive consequences for student achievement.

Importantly, more than 90 percent of the turnover of low-performing teachers occurs in high-poverty schools, which constitute 75 percent of all schools. The proportion of exiting teachers who are low performers is twice as high in these high-poverty schools as in low-poverty schools. In comparison with almost any other intervention, these are very large improvements for some of the school district's neediest students.

We do not claim that IMPACT caused all of the teacher turnover we observe. Although IMPACT did result in some teachers being dismissed, some attrition from teaching in DCPS as the system was implemented was surely voluntary. Nor do we know whether our results showing that teacher turnover benefited students in tested grades and subjects generalize to turnover for other teachers.

Policymakers looking to build on the DCPS model should also understand the implementation challenges and the potential for errors. Any teacher-evaluation system will make some number of objectionable mistakes in assigning ratings and consequences, and minimizing those instances will require expensive, sophisticated tools. Policymakers must weigh these costs against the substantial educational and economic benefits such systems can create for successive cohorts of students, both through avoiding the career-long retention of the lowest-performing teachers and through broad increases in performance in the overall teaching workforce.

More than 90 percent of the turnover of low-performing teachers occurs in high-poverty schools, which constitute 75 percent of all schools. The proportion of exiting teachers who are low performers is twice as high in these high-poverty schools as in low-poverty schools.

They should also expect continual change and improvement. DCPS has made modifications to IMPACT nearly every year, such as reducing the number of teacher observations, altering access to bonus and base-pay increases toward high-poverty schools, and reducing the weight of value-added performance measures. Most recently, the district announced it would rely on school principals for all observations and incorporate student-survey results in teachers' evaluations. DCPS also introduced LEAP, an intensive professional development program to help teachers improve their skills.

The challenge of improving the composition of teachers in DCPS is increasing. As the least-effective teachers exit, there are fewer such teachers who will leave over time. Whether DCPS can reap further performance benefits from compositional change in its workforce as it increases performance standards remains to be seen.

Regardless, our results indicate that, under a robust system of performance evaluation, the turnover of teachers can generate meaningful gains in student outcomes, particularly for the most disadvantaged students. The eight-year history of IMPACT shows that such efforts may incur political consequences, but are not politically impossible.

Thomas Dee is a professor of education and director of the Center for Education Policy Analysis at Stanford University, and a research associate at the National Bureau of Economic Research. James Wyckoff is the Curry Memorial Professor of Education and Policy and director of EdPolicyWorks at the University of Virginia. This article is based on research published in the March 2017 issue of Educational Evaluation and Policy Analysis *and in the March 2015 issue of the* Journal of Policy Analysis and Management.