# IS EFFECTIVE TEACHER EVALUATION SUSTAINABLE? EVIDENCE FROM DISTRICT OF COLUMBIA PUBLIC SCHOOLS

**Thomas S. Dee**
Graduate School of Education
Stanford University
Stanford, CA 94305-3001
tdee@stanford.edu

**Jessalynn James**
Annenberg Institute
Brown University
Providence, RI 02912
jessalynn_james@brown.edu

**Jim Wyckoff**
(corresponding author)
Curry School of Education
University of Virginia
Charlottesville, VA 22904-4277
wyckoff@virginia.edu

## Abstract

Ten years ago, many policy makers viewed the reform of teacher evaluation as a highly promising mechanism to improve teacher effectiveness and student achievement. Recently, that enthusiasm has dimmed as the available evidence suggests the subsequent reforms had a mixed record of implementation and efficacy. Even in districts where there was evidence of efficacy, the early promise of teacher evaluation may not be sustainable as these systems mature and change. This study examines the evolving design of IMPACT, the teacher evaluation system in the District of Columbia Public Schools. We describe the recent changes to IMPACT, which include higher performance standards for lower-performing teachers and a reduced emphasis on value-added test scores. Descriptive evidence on the dynamics of teacher retention and performance under this redesigned system indicates that lower-performing teachers are particularly likely to either leave or improve. Corresponding causal evidence similarly indicates that imminent dismissal threats for persistently low-performing teachers increased both teacher attrition and the performance of returning teachers. These findings suggest that teacher evaluation can provide a sustained mechanism for improving the quality of teaching.

https://doi.org/10.1162/edfp_a_00303

## 1. INTRODUCTION

Ten years ago, many education reformers championed rigorous and consequential teacher evaluation as an intervention that would improve the effectiveness of the teacher workforce and, in turn, increase student outcomes. In particular, both the federal government and prominent philanthropies encouraged such reforms through a variety of high-profile initiatives (e.g., Race to the Top, Teacher Incentive Fund, the Measures of Effective Teaching Project, No Child Left Behind waivers, and Intensive Partnerships for Effective Teaching). In response, most states and school districts designed and implemented new teacher evaluation systems (Steinberg and Donaldson 2016).

As reports on the effects of these teacher evaluation reforms have begun to accumulate, the corresponding public discussion has arguably become muddled. At a high level, states and school districts designed very similar systems. They all contained a teacher observation component and most included some form of student achievement outcomes for which the teacher is responsible (Steinberg and Donaldson 2016; Kraft and Gilmour 2017; Walsh et al. 2017). Some evidence suggests that rigorous teacher evaluation improved teaching and student outcomes in Washington, DC (Dee and Wyckoff 2015; Adnot et al. 2017), Chicago (Steinberg and Sartain 2015), Cincinnati (Taylor and Tyler 2012), and Houston (Cullen, Koedel, and Parsons 2017). Nonetheless, there is a growing public narrative that teacher evaluation reform has been a costly failure (Strauss 2015; Gates and Gates 2018; Iasevoli 2018) and a waste of resources (Dynarski 2016; Walsh et al. 2017). For example, a recent RAND study (Stecher et al. 2018) of three school districts and four charter management organizations found that teacher evaluation did not improve student achievement, but also suffered from "incomplete implementation" (p. 2).

The logistic and political challenges to implementing meaningful and informative teacher evaluation appear to be widespread. Kraft and Gilmour (2017) surveyed twenty-four states with teacher evaluation reforms and found, in most states, roughly 95 percent of teachers are still rated as effective or better. This finding is strikingly similar to those reported in *The Widget Effect*, a report from The New Teacher Project that precipitated much of the discussion regarding teacher evaluation reform (Weisberg et al. 2009). Currently, we know relatively little about why the implementation of teacher evaluation practices differs across contexts. And, more generally, we know relatively little about whether and under what circumstances teacher evaluation reforms have produced systematic changes in teaching and learning.

Even if teacher evaluation reforms produced meaningful *early* effects during the surge of enthusiasm and initial focus, the implementation literature offers ample cautions that such effects might not be maintained (Fixsen et al. 2005). Unless reforms altered school-level organizational cultures, effectively creating buy-in from principals and teachers, the forces that maintained the status quo pre-reform are likely to diminish the effects of these efforts. From this perspective, teacher evaluation is particularly vulnerable. The catalysts for teacher evaluation initiatives were typically top–down in nature and the design and implementation of teacher evaluation was often hurried to meet federal grant eligibility deadlines. Moreover, implementation often minimized or ignored the concerns of principals, teachers, and teacher unions (Chuong 2014; McNeil

2014). To become sustainable—the implementation literature suggests—such reforms would need to be implemented robustly and adapted over time to feedback and changing circumstances. Administrators need to provide continuing support and leadership; teachers and principals must find teacher evaluations practical and useful (Fixsen et al. 2005).

It is against this backdrop that we provide new evidence on IMPACT, the controversial teacher evaluation system in the District of Columbia Public Schools (DCPS). Prior research has documented that aspects of IMPACT initially improved teacher performance (Dee and Wyckoff 2015) and student achievement (Adnot et al. 2017). In this paper, we examine the evolving design features of IMPACT, the associated descriptive changes in the teacher workforce, and the corresponding causal effects of incentives on teacher attrition and performance under this mature and redesigned system. Notably, the design changes to IMPACT include a deemphasis on evaluating teachers with conventional value-added test scores and an increase in the performance standards. The higher expectations for teacher performance include a new rating category (i.e., "Developing") for lower-performing teachers who would have previously been considered "Effective." Even in the absence of these design changes, the longer-term effects of IMPACT's incentives are an open empirical question. For example, these reforms might be sustained if they remained well-implemented and if they catalyzed positive changes in school culture and performance. Alternatively, their effects might be attenuated in the context of an improving teacher workforce, as well as in response to the presence of leadership turnover, shifts in organizational focus, and internal pressure to limit their most binding consequences.

We begin by describing the key design features and their evolution into the "IMPACT 3.0" system, which was in place beginning with the 2012–13 school year. We then examine descriptively the dynamics of teacher retention and performance under IMPACT during the period from 2012–13 to 2015–16. Overall, we find lower-performing teachers are substantially more likely to either leave DCPS or to improve their performance relative to higher-performing teachers. We also provide corresponding causal evidence on this relationship through a regression discontinuity (RD) design that focuses on IMPACT's high-powered dismissal threat. Specifically, we examine the effects on teacher retention and performance of being rated as "Minimally Effective" (ME) instead of "Developing" (D), or D rather than "Effective" (E). The ME/D treatment contrast effectively compares the credible and immediate dismissal threat for ME teachers who do not improve immediately to the incentives faced by D-rated teachers who instead have two years to achieve an E rating. The D/E treatment contrast compares the incentives to improve within two years to teachers who receive no sanctions. IMPACT also provides incentives for teachers to score "Highly Effective" (HE); however, the changes to IMPACT, which we describe in detail below, made the incentive contrast at the E/HE threshold difficult to analyze. Consistent with the descriptive evidence, we find that facing a performance-based dismissal threat increased the voluntary attrition of lower-performing teachers. We also find qualified evidence that such threats increased the performance of teachers who returned. Our study concludes with a discussion of the implications of these findings for teacher evaluation research and policy.

## 2. INCENTIVES AND EVALUATION IN WASHINGTON, DC

In 2007, following his election on a reformist agenda, Mayor Adrian Fenty secured approval for mayoral control of DCPS. The low-income, largely minority district suffered from chronically low academic achievement and persistently struggled to make meaningful improvements. For example, DCPS's scores on the National Assessment of Educational Progress math tests in 2007 were lower than any other state or participating urban district in the country. The District was also among the lowest in reading performance (NAEP 2007). Before long, the quality of DCPS's teaching force became a focal point for these reforms. Evidence of the importance of teachers for driving student outcomes (e.g., Rockoff 2004; Rivkin, Hanushek, and Kain 2005; Gordon, Kane, and Staiger 2006) provided a motivation for this focus. Students in high-poverty schools are the least likely to have high-quality teachers, and poor schools attract less-experienced teachers and have higher rates of teacher attrition (Clotfelter, Ladd, and Vigdor 2005). Additionally, evidence suggests that the largest impacts of teacher quality occur for less-advantaged students, specifically African American students and those whose performance is in the low and middle ranges of the achievement distribution (Aaronson, Barrow, and Sander 2007).

It was in this context that, in 2009, under the direction of then-Chancellor Michelle Rhee, DCPS implemented IMPACT, a teacher-performance-assessment system. For an insightful discussion of the design and implementation of IMPACT, see Toch (2018). A fundamental intent of IMPACT was to incent and reward high-quality teaching, while removing low-performing teachers who failed to make adequate improvements. In the 2012–13 school year, DCPS changed several design features of IMPACT. Four features define much of IMPACT's structure: (1) the components of the multi-measure evaluation system; (2) the rating categories that distinguish teacher performance levels; (3) the thresholds that determine rating categories; and (4) the stakes associated with rating categories. Each of these has changed since IMPACT's inception to address feedback from teachers and evolving goals for improving student performance. Taken together, these changes became known in the district as IMPACT 3.0.

### Multi-measure Components

The components that make up teachers' IMPACT scores, and their weighting (table 1), depend on the grades and subjects taught. The majority of general-education teachers (80 percent) teach in grades and subjects for which value-added scores based on standardized tests cannot be defined. For these "Group 2" teachers, 75 percent of overall IMPACT scores are based on a classroom observation measure, the Teaching and Learning Framework (TLF). TLF scores reflect average performance across nine domains, measured as many as five times during the school year by a combination of in-school and external evaluators. Table 1 shows the evolving composition and weighting of these evaluation components. For teachers in tested grades and subjects (Group 1), the largest contributor to their IMPACT scores was based on student achievement, as measured by individual value-added scores (IVA). IVA was calculated utilizing a typical state achievement test, the DC-CAS, until 2014–15, when DCPS adopted the PARCC exam. For the first two years of the PARCC exam (2014–15 and 2015–16) IMPACT for Group 1 teachers did not include IVA over concerns that teachers needed time to adjust

**Table 1.** IMPACT Score Components 2009–10 through 2015–16

| | IMPACT 1.0–2.0 | | IMPACT 3.0 | | |
|---|---|---|---|---|---|
| | 2009–10 to 2011–12 | | 2012–13 to 2013–14 | | 2014–15 to 2015–16 |
| IMPACT Components | Group 1 | Group 2 | Group 1 | Group 2 | Groups 1 and 2 |
| IVA | 50% | 0% | 35% | 0% | 0% |
| TLF | 35% | 75% | 40% | 75% | 75% |
| TAS | 0% | 10% | 15% | 15% | 15% |
| CSC | 10% | 10% | 10% | 10% | 10% |
| School value-added | 5% | 5% | 0% | 0% | 0% |

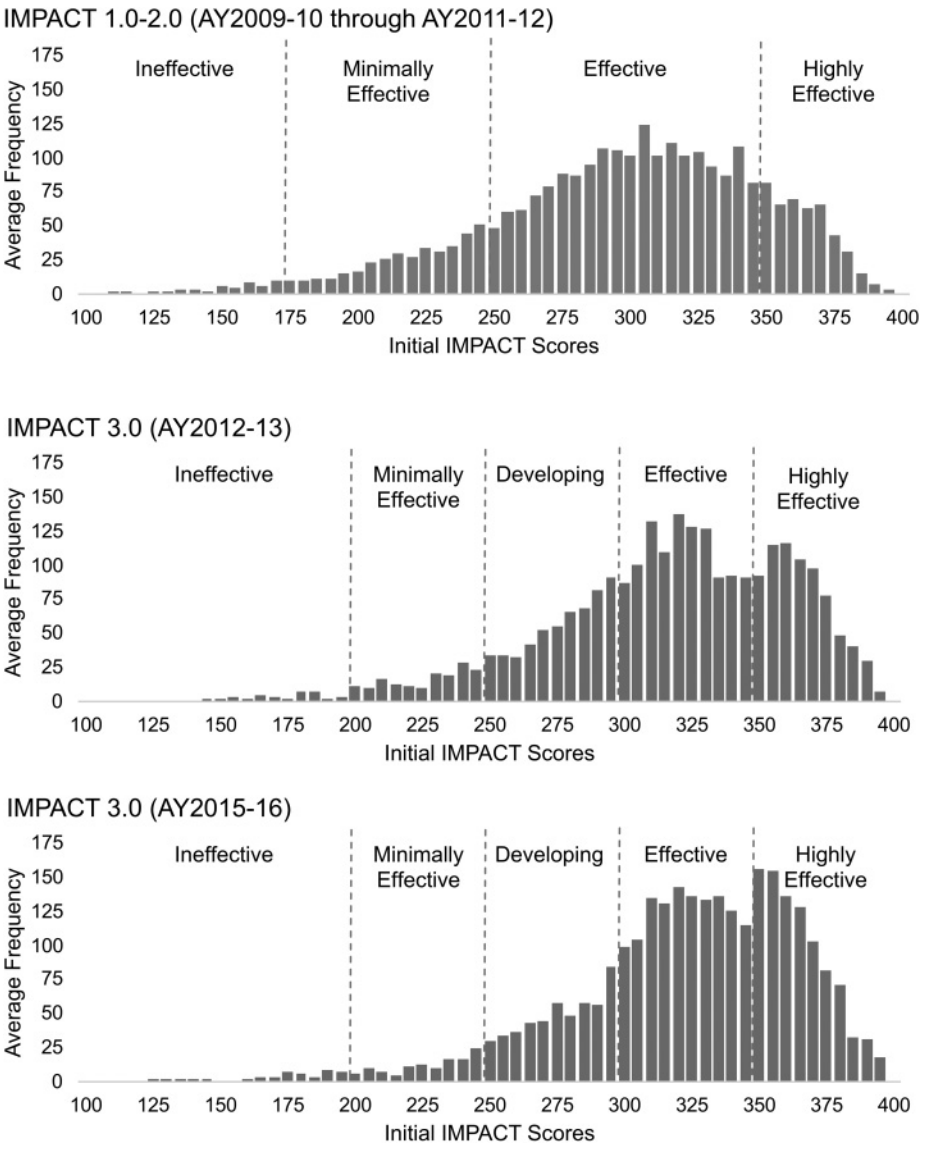
*Notes:* Group 1 consists only of those reading and mathematics teachers in grades for which it is possible to define value added with the available assessment data. IMPACT scores can also be adjusted downward for "Core Professionalism" violations reported by principals. Group 1 teachers did not have Individual Value Added (IVA) calculated during the first two years of the PARCC exam (academic years 2015 and 2016); in those years, Group 1 teachers had the same score components and weights as Group 2 teachers. The Commitment to the School Community (CSC) measure is a rubric-based assessment, scored by the school principal, of the teacher's contributions to the professional life of the school. The Teacher-Assessed Student Achievement Data (TAS) component is a measure of student performance on a teacher-selected assessment, where performance is evaluated relative to targets set at the start of the school year; the school principal must approve both the selected measure and the teacher-developed goals. TLF = Teaching and Learning Framework.

to PARCC. In addition, a small weight was applied to teachers' Commitment to the School Community measure (CSC), a rubric-based assessment, scored by the school principal, of the teacher's contributions to the professional life of the school. Group 2 (general education) teachers, for whom value-added scores were not available, were also evaluated according to Teacher-Assessed Student Achievement Data (TAS)—a measure of student performance on a teacher-selected assessment, where performance is evaluated relative to targets set at the start of the school year; the school principal must approve both the selected TAS measure and the corresponding goals.

Under IMPACT 3.0, the weights applied to these components changed substantially. In particular, the emphasis put on test-based value-added measures fell. DCPS eliminated school-level value added entirely in response to teachers' concerns that they had virtually no control over their scores on this school-level measure. In addition, the test data upon which Group 1 teachers were evaluated were not solely IVA on the standardized assessment; Group 1 teachers, following the IMPACT 3.0 reforms, were in part evaluated on self-selected student achievement measures (i.e., TAS). The stated intent of these changes was to reduce anxiety for Group 1 teachers, who expressed concern that such a large part of their IMPACT score was based on high-stakes value-added measures.

### Teacher Performance Categories

During its first three years, teachers were assigned to one of four rating categories—Highly Effective (HE), Effective (E), Minimally Effective (ME), and Ineffective (I)—based on their overall IMPACT score, which ranged from 100 to 400. In academic year (AY) 2012–13, DCPS created a new performance category—Developing (D)—by dividing the Effective category in half, with the lower portion becoming the Developing




*Notes:* IMPACT scores reported here are initial scores, assigned prior to the appeals process. Very few appeals result in revised scores. Sample consists of general education teachers in District of Columbia Public Schools. The distribution of scores around the Effective/Highly Effective threshold may indicate potential manipulation of scores; while it is possible manipulation occurs at this point in the distribution given that teachers with consistently high performance are subject to fewer classroom observations and can therefore see their overall scores more easily changed by a single classroom observation, this threshold is not one we focus on in this paper. AY = academic year.

**Figure 1.** Distribution of IMPACT Scores by Year and Rating

category. The motivation for this change included evidence that the prior Effective range reflected considerable variability in teacher performance and a desire to signal increased urgency to improve teaching skills and student outcomes. Initial and revised thresholds are shown in figure 1 and table 2. The intent of the increased

**Table 2.** IMPACT Ratings, Separation, and Extra Compensation Criteria, 2009–10 to 2014–15

| Category | | 2009–10 to 2011–12 | 2012–13 to 2014–15 |
|---|---|---|---|
| Scoring bands for performance ratings | | 100–174: Ineffective (I)<br>175–249: Minimally Effective (ME)<br>250–349: Effective (E)<br>350–400: Highly Effective (HE) | 100–199: I<br>200–249: ME<br>250–299: Developing (D)<br>300–349: E<br>350–400: HE |
| Separation criteria | | Separation after 1 I rating, or 2 consecutive ME ratings | Separation after 1 I rating, 2 consecutive ME ratings, 1 D followed by 1 ME rating, or 3 consecutive ratings below E |
| **Compensation** | | | |
| Bonus pay | **Eligibility** | **Teachers in all schools scoring HE** | **Teachers in all schools scoring HE** |
| | FRPL ≥ 60% | $10,000, plus $10,000 for teachers in Group 1, plus $5,000 for teachers in high-need subject | $10,000, plus $5,000 for teachers in Group 1, plus $10,000 for teachers in 40 lowest-performing schools |
| | FRPL < 60% | $5,000, plus $5,000 for teachers in Group 1, plus $2,500 for teachers in high-need subject | $2,000, plus $1,000 for teachers with value-added |
| Base pay increase | **Eligibility** | **Teachers in all schools** | **Teachers in schools with ≥60% FRPL** |
| | FRPL >= 60% | 2 consecutive years of HE ratings = Master's band + 5-year service credit | Advanced teacher: 2-year service credit<br>Distinguished teacher: Master's band + 5-year service credit<br>Expert teacher: PhD band + 5-year service credit |
| | FRPL < 60% | 2 consecutive years of HE ratings = Master's band + 3-year service credit | None |

*Notes:* Teachers must be "teaching in a high-poverty school during the year in which you qualify for a service credit, and during the following school year" in order to be eligible for the base salary increase (LIFT guidebook, 2012–13, p. 18). FRPL = free or reduced-price lunch.

performance standards embedded in these threshold changes was to encourage teachers to strengthen their teaching skills.

### Performance Stakes

Teachers identified as I by IMPACT have always faced dismissal at the end of the school year in which the rating was earned, as have teachers who scored twice consecutively as ME (table 2). Similarly, teachers rated HE received substantial one-time bonus payments, with amounts varying by the subject and grade level taught and the proportion of students in the teachers' schools receiving free and reduced-price lunch. In addition, before IMPACT 3.0, teachers who attained an HE rating for two consecutive years were eligible to receive a considerable base pay increase. The bonus and base-pay increases varied depending on whether teachers were teaching a subject with value-added scores, were teaching in high-poverty schools, and/or were teaching a high-need subject (table 2).

Beginning in AY 2012–13, IMPACT 3.0 modified the stakes associated with different rating categories. As before, teachers would be dismissed with one I or two consecutive ME ratings. However, with the introduction of the D category, teachers would be separated with three consecutive D ratings (or one D and a subsequent ME or I rating). DCPS also introduced a performance-based career ladder for teachers: The Leadership

Initiative for Teachers (LIFT). LIFT was intended to provide teachers with additional recognition and professional opportunities.[1] Importantly, LIFT also became the mechanism by which teachers' base-pay increases were determined. These base-pay increases became a function of the level and persistence of performance measured by IMPACT. The incentives for HE teachers also differ somewhat from those offered under the prior design of IMPACT (table 2). DCPS altered bonuses to create stronger incentives to teach in the forty most demanding schools in DCPS and substantially reduced incentives for teachers in low-poverty schools (i.e., those with less than 60 percent of students eligible for free and reduced-price lunch). These changes in stakes instantiated a focus on attracting and retaining HE teachers in high-poverty schools.

These design changes and the ongoing evolution of DCPS teachers coincided with changes in the distribution of teacher effectiveness, as shown by the graphs in figure 1. As is evident, the measured performance of teachers has meaningfully increased over time. For example, between 2009–10 and 2015–16, the median IMPACT score increased from 303 to 332 (i.e., a gain equivalent to 0.58 standard deviation [SD]). Before we examine teacher retention and performance under IMPACT 3.0, we address concerns recently raised about the manipulation of measured student outcomes in DCPS.

In general, the intended goals of accountability reforms in education are to provide teachers and school leaders with actionable information that can guide their improvement as well as with incentives that encourage those changes. IMPACT seeks to improve the effectiveness of the teaching workforce through the improvement of teaching skills and the attrition of teachers with unacceptably poor performance, and has adopted dismissal policies toward that end. A notable concern with output-based reforms is they may also cause some individuals to engage in unintended, counterproductive (and, in some cases, illegal) activities. For example, DCPS has recently come under scrutiny for inappropriately graduating students who had not met graduation requirements in an effort to improve graduation rates, a widely cited measure of educational success, and one that can play a small role in DCPS principal evaluations. School leaders were also caught manipulating—or pressuring their teachers to manipulate—student attendance and course credit data to meet school-level performance targets (Balingit and Tran 2018; Brown, Strauss, and Stein 2018; McGee 2018).

These allegations, although notable and troubling, are not directly salient for IMPACT. Graduation rates, attendance rates, and credit accumulation are not a component of teachers' IMPACT scores. Instead, IMPACT heavily weights classroom observations intended to induce teachers to improve diverse pedagogical skills and behaviors. In theory, the emphasis on TLF could encourage manipulation by principals who want to support teachers' ratings. However, the presence of additional TLF ratings by external evaluators, who typically conduct 40 percent of observations, and the corresponding system of principal accountability, suggest that such manipulation is unlikely for all but the most effective teachers.[2] We are aware of no assertions of manipulation to improve

1. The opportunities associated with advancing through LIFT stages include developing curricular materials, mentoring colleagues, and being eligible for certain fellowship opportunities. More information about the LIFT program is available on the DCPS Web site at https://dcps.dc.gov/page/leadership-initiative-teachers-lift.
2. The variability in principals' TLF ratings is also inconsistent with widespread manipulation. Dee and Wyckoff (2015) also find that IMPACT incentives generated similar increases in the TLF ratings by principals and external evaluators.

teacher IMPACT ratings.[3] Though manipulation of IMPACT scores seems unlikely, we explicitly examine the density of observations near the relevant thresholds as part of our analysis and find no evidence of such manipulation.

In sum, IMPACT 3.0 signals the intent by DCPS to make additional improvements in student academic performance by increasing the performance of teachers. Under IMPACT 2.0, about 70 percent of teachers earned an E rating and this performance range was quite broad. Creating the D category by dividing the E range in half and broadening the range for I-rated teachers sent a strong signal that DCPS believed they could meaningfully improve teacher effectiveness. DCPS also signaled an intent to increasingly focus on its lowest-performing schools. Financial incentives for high-performing teachers were dramatically reduced in low-poverty schools, where base-pay incentives were eliminated and bonuses for high performance cut by 75 percent. IMPACT 3.0 also included important elements to address concerns raised by teachers. The weight applied to IVA was reduced from 50 to 35 percent for applicable teachers. The elimination of school-level value added also addressed a long-standing concern by teachers that this measure was beyond their direct control. The career ladder, LIFT, added formal recognition and rewards to teachers as they realized professional development milestones.

## 3. LITERATURE REVIEW

The conceptual foundations for teacher evaluation policies focus on two broad mechanisms. One mechanism involves how incentives may shape the development and performance of extant teachers in ways that are beneficial to students. For example, programs that provide teachers with clear and actionable feedback on the character of their classroom performance can provide targeted support to their professional development. The presence of sanctions or rewards based on their performance can also encourage teachers both to increase their effort and to reallocate their instructional focus toward effective practices.

The empirical literature examining the effects of performance assessment and incentives on teacher performance is mixed. In particular, several small-scale and experimental attempts to use financial incentives to improve teachers' performance find limited or null effects (Springer et al. 2010; Marsh et al. 2011; Springer et al. 2012; Fryer 2013).[4] However, there are some studies in which teachers have responded to such incentives with improved performance (e.g., Balch and Springer 2015; Chiang et al. 2017). Furthermore, some studies (e.g., Taylor and Tyler 2012; Steinberg and Sartain 2015) provide evidence that evaluations do not necessarily need to be linked to rewards or sanctions to enhance teachers' practice. A potentially important unintended consequence is that high-stakes evaluations might encourage unintended behaviors such as cheating, particularly when a single performance outcome is emphasized (Apperson, Bueno, and Sass 2016). Although such responses have been observed where stakes are

3. Allegations of cheating on the high-stakes test in DCPS received extensive coverage in the press prior to 2012–13; we are unaware of any allegations since. Dee and Wyckoff (2015) address the allegations of cheating for this earlier period and find cheating was very limited and had no effect on their estimates of the effect of IMPACT.

4. The incentives examined in these studies may be weak for a variety of reasons: low dollar amounts, group rather than individual incentives, a focus on cash for test scores rather than more direct measures of teacher performance, and the expectation that the incentives are temporary rather than an enduring policy change.

tied to school- or student-level performance (e.g., Jacob and Levitt 2003; Dee et al. 2019), we do not know of such evidence in the context of teacher-level accountability systems.

The second mechanism that motivates teacher evaluation reforms concerns the composition of the teacher workforce—that is, the expectation that they will increase the recruitment and retention of high-performing teachers while also encouraging the attrition of low-performing teachers (Goldhaber 2015). While the evidence linking incentives to retention is by no means universally positive, incentive policies have generally been associated with improved retention. Fulbeck (2014), for example, found that Denver Public School District's ProComp program, which awards additional financial compensation for a variety of performance criteria, extra credentials, and teaching in high-poverty schools, is associated with significantly improved teacher retention within a school, though these retention effects are substantially smaller for high-poverty schools. North Carolina had similar success with a briefly implemented program that awarded bonuses to teachers of high-need subjects who taught in low-income and low-performing schools (Clotfelter et al. 2008). Chicago's Teacher Advancement Program, which awarded bonuses according to value added and classroom observation scores, as well as to teachers who took on leadership and mentorship roles within their schools, was also associated with improved school-level retention (Glazerman and Seifullah 2012). In Tennessee, teachers in low-performing schools who earned performance bonuses were more likely to be retained than their peers who scored just below the threshold of bonus eligibility, but this effect was concentrated only among teachers in tested grades and subjects (Springer, Swain, and Rodriguez 2016).

Incentives and evaluation can also influence teaching composition by encouraging higher-performing teachers to enter the profession. Such effects are less well documented in the literature, but simulations of incentive-based evaluation on entry into the teacher labor market (Rothstein 2015) suggest that performance-based contracts can alter the performance distribution of the teaching workforce by enticing higher-ability teachers while disincentivizing the entry or retention of lower-ability teachers. These effects, however, may be extremely small, given that those who are new to teaching generally have little confirmation of their performance ability from which to assess their probability of earning incentives. The most compelling evidence of selection-into-teaching effects comes from California, which briefly offered a $20,000 Governor's Teaching Fellowship to the most competitive students from accredited postbaccalaureate teacher licensure programs in return for teaching in low-performing schools. Steele, Murnane, and Willett (2010) found that these novice teachers were significantly more likely to begin their teaching careers in low-performing schools than they would have in the absence of the Fellowship program.

In general, few studies have examined the extent to which teacher evaluation reforms produce shifts in the quality and composition of the teaching force *as well as* ensuing effects on student achievement. Many of those who have looked at both teacher- and student-level outcomes have reached only limited conclusions. Teacher incentive programs may more feasibly shift immediate outcomes, such as teacher's retention and practice, than more distal outcomes such as student achievement; indeed, with other interventions, effects are larger with more closely aligned outcomes (Ruiz-Primo et al. 2002; Kraft 2020). While some incentivized outcomes can provide formative feedback about teaching quality (e.g., classroom observations) from which teachers can

glean information about how they might improve their performance, other incentivized outcomes—most notably those based on student achievement—are purely summative and do not come with embedded prescriptions for improvement that might in turn improve student outcomes. These mechanisms may be more effective when accompanied by feedback that is specifically aligned to teachers' performance on measures of their practice (e.g., Taylor and Tyler 2012; Steinberg and Sartain 2015; Kane et al. 2020). Effects on both practice and retention might also need to be substantial in size in order to produce measurable improvements to student outcomes (Cullen, Koedel, and Parsons 2021).

#### Evidence from DCPS

This prior literature provides an important context for understanding the mechanisms through which IMPACT might improve DCPS's teaching quality (i.e., performance, recruitment, retention, and attrition). Recent empirical studies based on the earliest years of IMPACT suggest that DCPS's reforms had positive impact on most of these fronts.[5] For example, there is evidence that IMPACT influenced the composition of the DCPS teaching workforce in a manner that improved teacher effectiveness and student achievement. Using a regression discontinuity design, Dee and Wyckoff (2015) found that a dismissal threat for low-performing teachers led to a 50 percent increase in the attrition of those teachers, indicating that the program successfully induces voluntary departure of its weaker teachers. Such teacher turnover could actually harm student learning through the disruption of teacher teams and through hiring less-qualified teachers. However, Adnot et al. (2017) find that performance-based dismissals and attrition in DCPS led to replacements who were substantially more effective at raising student achievement. These achievement effects were particularly strong for students in high-poverty schools.

The early effects of IMPACT were not purely compositional, however. Dee and Wyckoff (2015) also examined the effect of strong incentive contrasts at consequential performance thresholds on retained teachers' next-year performance. They found positive performance effects for high-performing teachers facing potentially large financial rewards, as well as for low-performing teachers who faced potential dismissal but remained teaching in DCPS. Among those who returned teaching the next year, both ME and HE teachers improved by approximately 25 percent of a standard deviation of IMPACT points. Importantly, Dee and Wyckoff also found that ME teachers' performance effects were in part driven by improvements to their value-added scores, suggesting that incentivized teachers improved in ways that extended to student learning.

In summary, the high-fidelity implementation and sustained impact of large-scale educational reforms have proven difficult to achieve (Fixsen et al. 2005; Chiang et al. 2017; Stecher et al. 2018). Indeed, as described above, the evidence from rigorous assessments of teacher evaluation is mixed, raising important questions regarding the sustainability of this reform even in the contexts where it met with initial success. We turn to an examination of whether IMPACT was able to sustain the initial substantial

5. The one exception is teacher recruitment and selection into DCPS. We know little about the causal effects of IMPACT because the policy went to scale simultaneously. However, Jacob et al. (2018) examine the screening of DCPS teacher applicants under IMPACT. Their description indicates that, under IMPACT, DCPS has a larger number of teacher applicants and a multifaceted screening process than exists in most districts.

**Table 3.** Mean Characteristics of Analytic Samples

| | Minimally Effective (ME) / Developing (D) | Developing (D) / Effective (E) |
|---|---|---|
| Retention next year | 0.75 | 0.83 |
| Next-year IMPACT score | 297 | 321 |
| Initial IMPACT score | 269 | 311 |
| Group 1 | 0.25 | 0.24 |
| Female | 0.72 | 0.70 |
| Gender missing | 0.01 | 0.01 |
| Black | 0.56 | 0.51 |
| White | 0.20 | 0.30 |
| Hispanic | 0.05 | 0.05 |
| Graduate degree | 0.62 | 0.65 |
| 0–3 years of experience | 0.32 | 0.31 |
| 4–9 years of experience | 0.30 | 0.32 |
| 10+ years of experience | 0.35 | 0.36 |
| AY 2012–13 | 0.35 | 0.37 |
| AY 2013–14 | 0.34 | 0.30 |
| AY 2014–15 | 0.31 | 0.32 |

*Notes:* The ME sample consists of 1,809 general-education teachers in the 2012–13 through 2014–15 academic years who received a ME or D rating and were not rated ME in the preceding year. The D sample consists of 4,105 general-education teachers in the 2012–13 through 2014–15 academic years who received a D or E rating and were not rated ME or D in the preceding year. See text for details. AY = academic year.

improvements in teacher effectiveness and student achievement both as the program matured and as its design evolved in important ways.

## 4. DATA AND SAMPLE

We base our analysis on a panel of teacher-level administrative data spanning from the start of IMPACT in AY 2009–10 through AY 2015–16. These data include, for all teachers in DCPS, information on teachers' IMPACT scores, ratings, and consequences, as well as demographic characteristics (e.g., race and gender), background (i.e., education and experience), and information about the schools in which they work and the students they teach (table 3). The IMPACT data include initial scores, as well as final scores that reflect the very small number of cases where scores were revised or successfully appealed. We use these data to create our two outcome variables: retention and next-year IMPACT score.

Our analysis focuses on what is arguably IMPACT's most potent incentive: the risk of dismissal for teachers who received an ME rating in the preceding year, as well as the less immediate risk of dismissal for teachers who received a D rating in the preceding year. We limit our analysis of incentives to the ME/D and D/E thresholds. Treatment at the E/HE threshold is variable and relies upon different criteria over time, and because the sample sizes are quite small across many of these treatment conditions, we do not explore treatment effects for high-performing teachers incentivized by bonus pay or salary increases.

The full sample consists of 17,465 teacher-by-year observations of teachers who received IMPACT ratings between AY 2010–11 and AY 2014–15, with approximately 3,500

teacher ratings per year. Of these observations, 13,192 (76 percent) are general education teachers—roughly 2,600 teachers per year. We use these data to create two distinct analytic datasets: one for teachers at the ME/D threshold, and a second for teachers at the D/E threshold.

To create our ME analytic datasets, we construct samples that include general education teachers whose rating in year *t* places them on either side of the ME/E cutoff in IMPACT 2.0 (AY 2010–11 to 2011–12) and the ME/D cutoff in IMPACT 3.0 (AY 2012–13 to 2014–15). In both cases, teachers who are rated ME face involuntary separation if they receive a second consecutive ME rating. This reduces our first analytic sample to 4,300 teachers in IMPACT 2.0 and 1,980 teachers in IMPACT 3.0. We omit teachers from IMPACT 1.0 from our analysis because of anecdotal evidence that teachers initially did not expect IMPACT to persist beyond its first year, which is further supported by null results in Dee and Wyckoff's (2015) analysis of IMPACT's initial years.

Teachers are assigned to the ME treatment group if their score (pre-appeals) placed them in the ME score range. Under IMPACT 2.0, ME scores ranged from 175 through 249, and under IMPACT 3.0 ME scores ranged from 200 through 249. Teachers who have scored their first ME rating must improve by the following year if they wish to retain their teaching positions. The teachers scoring at the next highest rating level do not face this threat. Before the 2012–13 changes, this was teachers earning an E rating (scoring between 250 and 349); following program revisions, this group consisted of teachers earning a D rating (those scoring between 250 and 299).

Any teachers not assigned to the ME treatment and the rating category just above it are removed from this analytic sample. To avoid conflation of voluntary and involuntary separation outcomes, the treatment sample is then restricted to teachers who did not have an ME or D rating in the prior year—ratings that result in involuntary dismissal when immediately followed by an ME rating. After these adjustments, the ME analytic sample consists of 3,888 teachers in IMPACT 2.0, 528 (14 percent) of whom are rated ME, and 1,809 teachers in IMPACT 3.0, of whom 370 (20 percent) are rated ME.

We create a second, distinct analytic dataset for estimating effects at the D/E threshold. We first restrict the overall sample to general education teachers whose rating in year *t* places them on either side of the D/E cutoff in IMPACT 3.0 (AY 2012–13 to 2014–15). Because the D rating category did not exist prior to IMPACT 3.0, there is no comparable IMPACT 2.0 dataset. Any teachers not assigned to the D treatment and the rating category just above it (E) are removed from the D analytic sample. This reduces our analytic sample to 3,996 teachers.

Teachers are assigned to the D treatment group if their pre-appeals score was between 250 and 299, placing them in the D score range. Teachers who have scored their first D rating must improve over the course of the next two years if they wish to retain their teaching positions, while those scoring just above D (E) face no dismissal threat. To avoid conflation of voluntary and involuntary separation outcomes, and to ensure a clean treatment contrast, the sample is then restricted to teachers who did not have an ME or D rating in the prior year—ratings that result in involuntary dismissal when immediately followed by an ME rating, or two consecutive ratings below E. The final D analytic sample consists of 3,271 teachers, 980 (30 percent) of whom are rated D.

We construct separate samples for these two treatment thresholds in part to avoid treatment overlap. The steps described above to create our samples (i.e., removing

**Table 4.** Reduced-Form Minimally Effective Intent-to-Treat Regression Discontinuity Estimates on Teacher Retention and Performance, by IMPACT Phase

| | Retention | | | | Next-Year IMPACT Score | | | |
|---|---|---|---|---|---|---|---|---|
| **Sample** | **(1)** | **(2)** | **(3)** | **(4)** | **(1)** | **(2)** | **(3)** | **(4)** |
| **IMPACT 2.0** | | | | | | | | |
| $I(S_{it} < 0)$ | −0.093* | −0.090* | −0.092* | −0.092 | 9.03+ | 8.01+ | 7.03 | 8.73 |
| | (0.046) | (0.044) | (0.042) | (0.062) | (4.93) | (4.80) | (4.43) | (6.41) |
| | *1,874* | *1,874* | *1,874* | *1,874* | *1,439* | *1,439* | *1,439* | *1,439* |
| Teacher controls | No | Yes | Yes | Yes | No | Yes | Yes | Yes |
| School fixed effects | No | No | Yes | Yes | No | No | Yes | Yes |
| Quadratic of running variable | No | No | No | Yes | No | No | No | Yes |
| AIC | 1,986 | 1,892 | 1,756 | 1,759 | 14,653 | 14,608 | 14,416 | 14,419 |
| **IMPACT 3.0** | | | | | | | | |
| $I(S_{it} < 0)$ | −0.117* | −0.104* | −0.114* | −0.138* | 12.89* | 12.19+ | 11.99* | 8.31 |
| | (0.052) | (0.050) | (0.047) | (0.069) | (6.52) | (6.45) | (6.04) | (9.09) |
| | *1,809* | *1,809* | *1,809* | *1,809* | *1,270* | *1,270* | *1,270* | *1,270* |
| Teacher controls | No | Yes | Yes | Yes | No | Yes | Yes | Yes |
| School fixed effects | No | No | Yes | Yes | No | No | Yes | Yes |
| Quadratic of running variable | No | No | No | Yes | No | No | No | Yes |
| AIC | 2,041 | 1,952 | 1,806 | 1,810 | 13,043 | 13,023 | 12,802 | 12,805 |

*Notes:* Robust standard errors are in parentheses and sample sizes are in *italics*. Models include year fixed effects and use uniform kernel weights. Treatment effects are estimated off of teachers who were not rated Minimally Effective in the prior year. Teacher covariates include gender, race, education, experience, and an indicator for whether the teacher is in a tested grade and subject (Group 1). We exclude academic year 2009–10 (IMPACT 1.0) because of evidence that IMPACT was not truly implemented at that point. AIC = Aikake information criterion.

$^{+}p < 0.1$; $^{*}p < 0.05$.

teachers with prior-year scores at the consequential threshold, relying on initial scores instead of final, post-appeal scores, and establishing fully separate analytic samples consisting only of teachers at and just above the given consequential threshold) ensure that we avoid complications associated with other incentive-relevant thresholds.

An additional important sampling distinction to note is that the cohorts we use to estimate IMPACT 2.0 results overlap with, but are not exactly the same as, those used in the Dee and Wyckoff (2015) study. We omit the first year of IMPACT (i.e., 2009–10) from the analysis because of anecdotal and empirical evidence—described in the 2015 study—that IMPACT was not truly implemented at that point, but include outcomes for an additional cohort of teachers (i.e., those evaluated in 2011–12), which was not yet available at the time of the earlier paper's publication. We also present in table 4 estimates from a narrower bandwidth (±50 IMPACT points) than that used in the 2015 study. We focus the paper on the IMPACT 3.0 period, given that most years of IMPACT 2.0 were covered in an earlier paper (Dee and Wyckoff 2015); however, for comparative purposes, in some cases we include estimates from pooled results across IMPACT 2.0. While this represents a slightly different set of IMPACT 2.0 cohorts than in the Dee and Wyckoff (2015) paper, estimates from each pair of pooled cohorts yields qualitatively similar IMPACT 2.0 effects on teachers' retention and performance.

## 5. METHODS

We first explore patterns in teachers' performance and retention descriptively by following teachers' retention decisions under IMPACT 3.0. We then turn to examining the

effects of IMPACT's dismissal threat on teacher retention and performance. Specifically, we rely on an RD design to estimate the effects of an ME or D rating. This approach effectively exploits the plausibly random variation in teachers' initial IMPACT ratings around the respective threshold to estimate local treatment effects. Our specifications take the following general form:

$$Y_{it} = \beta_0 + \delta(D_{it}) + f(S_{it}) + X_{it}\lambda + \tau_t + \varepsilon_{it}.$$

For each threshold, $Y_{it}$ represents teacher $i$'s retention or performance following year $t$ (as measured by next-year IMPACT scores); $\delta$ represents the effect of the teachers' IMPACT rating ($D_{it}$)—specifically, the effect of falling on the consequential side of the relevant cut point (i.e., scoring $\leq$249 for the ME/D threshold or $\leq$299 for the D/E threshold); $f(S_{it})$ is a flexible function of the assignment variable (i.e., the initial IMPACT score centered on the consequential threshold); $X_{it}$ is a vector of teacher covariates; $\tau_t$ represents year fixed effects to account for differences in the relationship between IMPACT assignment and baseline characteristics across years; and $\varepsilon_{it}$ is an individual- and year-specific error term. In addition, we also explore models of the RD that include school fixed effects. Given that teachers rated D have two additional years to attain a higher rating without immediate dismissal (in contrast to just one year for ME teachers), we also estimate effects on retention and performance in year $t + 2$ for the D analytic sample.

We use several methods to test the internal validity of our estimates following best practice for RD analyses (Lee and Lemieux 2009; WWC 2017; Cattaneo, Idrobo, and Titiunik 2019), including tests for robustness of results to assumptions about the functional form of the relationship between teachers' IMPACT scores and their retention or future performance. More specifically, our baseline specification controls for linear splines of the assignment variable above and below the respective threshold. However, we explore local linear regressions that use increasingly smaller bandwidths of scores around the consequential cut point. We also examine specifications that include higher-order polynomials of the assignment variable and that apply triangular kernel weights to regressions, such that greater weight is placed on scores closer to the threshold than those further away. These are discussed in our Results section and presented in the appendices to this paper (available in a separate online appendix that can be accessed on *Education Finance and Policy*'s Web site at https://doi.org/10.1162/edfp_a_00303).

In addition to functional form, a key assumption for RD analysis is the exogeneity of treatment. Nonrandom sorting of teachers to different score levels might be of particular concern given emerging evidence that some types of teachers (e.g., those from racial or ethnic minorities, or those serving disproportionately advantaged students) earn lower classroom observation scores, on average, than their peers (e.g., Drake, Auletto, and Cowen 2019). We test for such nonrandom assignment to treatment empirically, by estimating our regression specification with teachers' pretreatment characteristics on the left-hand side in lieu of retention and performance outcomes. If treatment at the threshold is randomly determined, we should find no significant effects on $\delta$ for any of these teacher covariates. Results from these regressions (table A.1) indicate no significant sorting of teachers to the ME treatment or control condition by observable characteristics at conventional significance levels; the probability of being assigned

to treatment for teachers with five through nine years of experience is significant at $\alpha = 0.10$. At the D threshold, our covariate balance tests suggest possible sorting of teachers by race and experience; white teachers and teachers with two to four years of experience are somewhat less likely to be rated D than E ($p < 0.05$). We observe no additional indication of potential covariate imbalance, and tests of the equality of coefficients indicate no statistical difference in rating assignment across teacher covariates. Regardless, we condition on these observable characteristics to limit potential endogeneity. Systematic score manipulation is quite unlikely in this context. This would be a concern, for example, if certain types of teachers were able to improve their initial scores to avoid assignment to the treatment, potentially confounding our treatment estimates. There are several reasons we believe this is not a concern in the case of IMPACT.

First, although it is conceivable that observation (TLF) scores could be manipulated if a school administrator were concerned about a teacher who faced separation based on prior-year IMPACT scores, giving that teacher a more generous TLF score as a result, this would be difficult to do in practice. While TLF scores are composed in part of ratings from administrators—who might manipulate scores given their contextual knowledge of teachers' performance and personalities—external Master Educators also rate teachers and would not be privy to information about a given teacher's prior performance. We explicitly test for this by comparing treatment estimates from our regression models (not shown) where the outcome is the principals' TLF score to models where the outcome is the TLF score assigned by Master Educators; the difference in treatment estimates by type of rater is statistically indistinguishable from zero. In addition, while observation measures make up a plurality of teachers' overall scores, those assigned by school administrators are only partial contributors to the overall evaluation score, contributing a typical weight of no more than 45 percent of total IMPACT scores, limiting principals' ability to precisely influence teachers' scores. DCPS also uses a scoring platform, *align*, to calibrate its raters; this makes it less likely that school-based evaluators would score lessons differently than the Master Educators.

Second, we use teachers' initial IMPACT scores, rather than the scores they may have received post-appeal. Doing so substantially mitigates against score manipulation and avoids violation of the exogeneity assumption. As shown in figure 2, there is a nearly sharp discontinuity in the probability of assignment to treatment for both the ME and D analytic samples, given a teacher's initial IMPACT score in AY 2012–13. When final, post-appeal IMPACT scores are used, there could be some manipulation occurring around the cut points, though potential effects of this manipulation are small, given that few teachers' IMPACT ratings are successfully appealed. In the 2012–13 through 2014–15 academic years, only fifty-six of the initial IMPACT ratings for Group 1 and Group 2 teachers across all of the ratings thresholds were changed following revisions or appeals, representing less than 1 percent of all ratings across the three years. Most of these appeals (82 percent) were granted in the first year of IMPACT 3.0, while the number of successful appeals granted in AY 2013–14 and AY 2014–15 declined respectively to one and nine. The use of initial, pre-appeal scores could diminish the external validity of findings; however, given that so few teachers succeed in their attempts at revising initial scores, any differences in findings would likely be negligible had there been no score revisions (or had the analysis been of treatment-on-treated, rather than intent-to-treat,

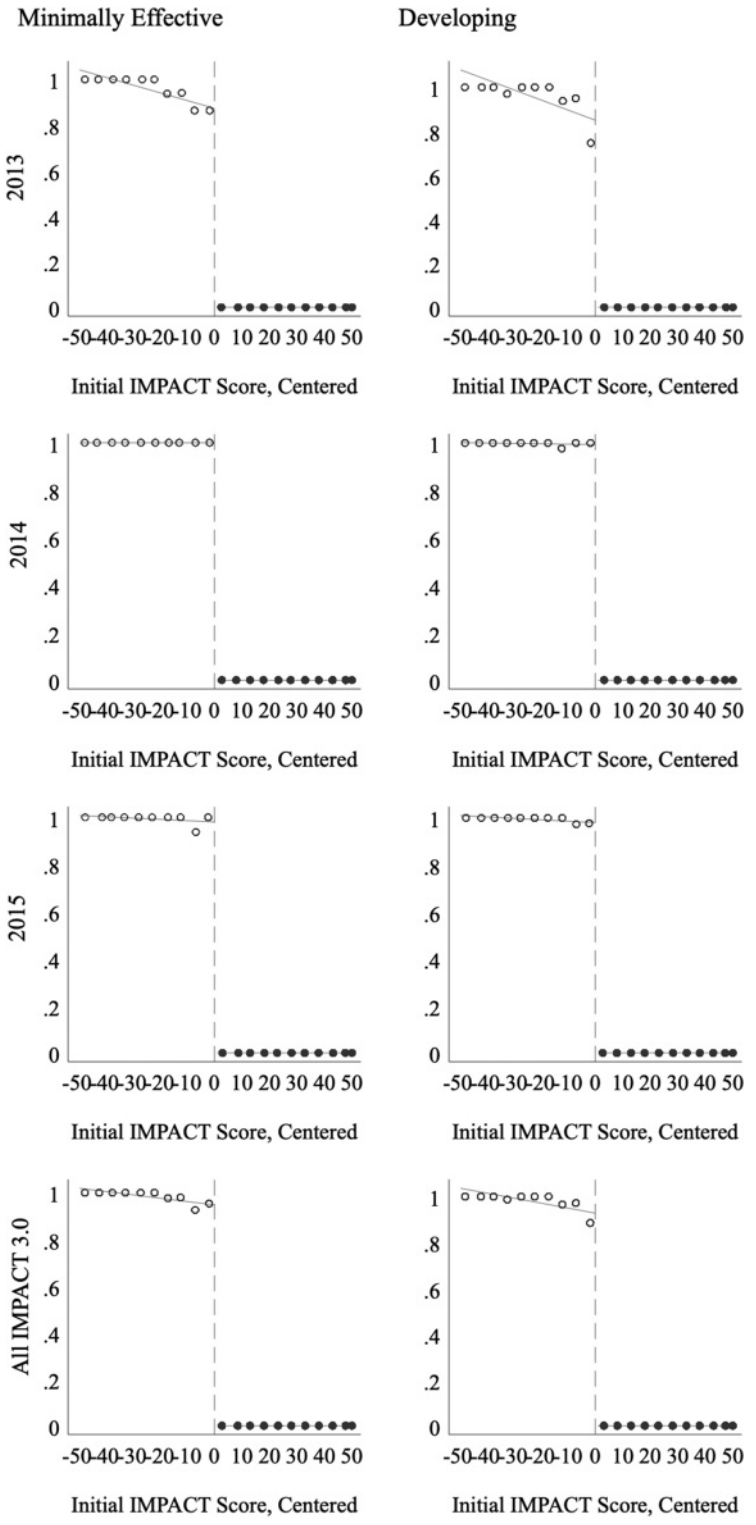



Figure 2. First Stage: Effect of Initial IMPACT Score on the Probability a Teacher Is Rated Minimally Effective or Developing at the Consequential Cutoff

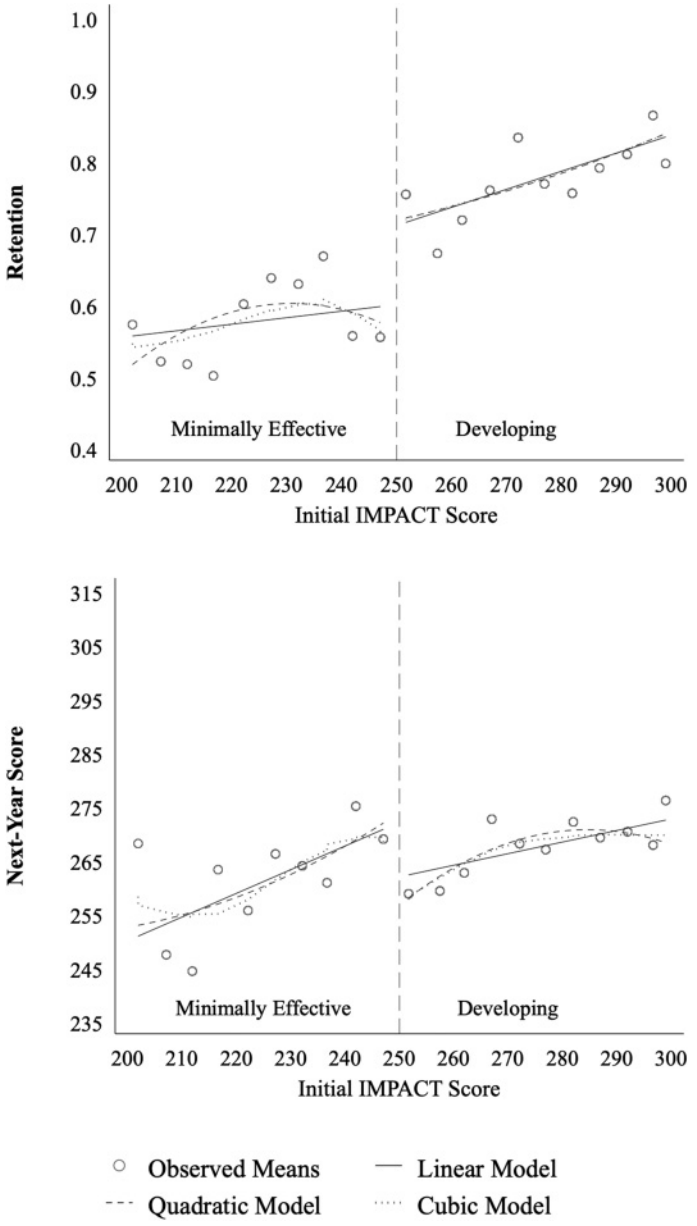
effects). In addition, fuzziness effects are largely isolated to AY 2012–13, following an error in the calculation of teachers' IVA scores.

Density tests of the distribution of observations through the ME and D thresholds provide direct empirical evidence that manipulation of the assignment variable did not occur (McCrary 2008). Specifically, we use the local-polynomial density estimators proposed by Cattaneo, Jansson, and Ma (2018, 2020) to test for discontinuity in the density of observations around the ME/D and D/E thresholds. This test relies on the assumption that if there were no systematic manipulation of scores around the threshold, we would observe continuous changes in the density of observations at the cutoff; conversely, evidence of discontinuous density at the threshold would suggest possible nonrandom sorting of teachers to ME or D ratings. We run this falsification test for each year of IMPACT 3.0 individually and for all three years in aggregate, finding no statistical difference in densities across the threshold within or across years. This evidence (figures A.1 and A.2 in the online appendix), further supports our assumption that treatment is exogenous at the ME/D and D/E thresholds.

Third, for an RD to be internally valid, an additional requirement is that the average outcome (in this case, either retention or next-year IMPACT scores) is a continuous function of teachers' current-year IMPACT scores, conditional on their IMPACT rating. Concerns about the violation of this assumption would be raised if the relationship between the two outcomes and teachers' IMPACT scores indicated discontinuities at points other than the consequential threshold. If there were no treatment effect, we would expect the relationship between initial IMPACT scores and retention or next-year performance to continue as is, without additional discontinuities beyond the consequential cut points. The graphs in figures 3 through 5 suggest that this assumption is not violated at the ME/D or D/E thresholds; however, because this relationship is noisy it is difficult to assess purely through visual evidence. To further test that this assumption is met, we run a series of RD models using "placebo" cut points. Assuming there is a discontinuity, or treatment effect, at the consequential threshold, there should be no other detectable effects at thresholds where we would not expect to see them. These placebo tests (available in table A.2 of the online appendix) produce no significant results at any point other than the cutoff between ME and D ratings for the ME analytic sample and the cutoff between D and E ratings for the D analytic sample.

Another potential threat to the validity of our estimates is the possibility of differential attrition from the sample across the threshold of analysis (WWC 2017). There are, however, two key reasons why attrition is not a concern in this context for teachers' retention. First, we assess intent-to-treat effects based on initial IMPACT score assignment, thereby defining treatment as the threat of dismissal associated with having initially scored at the ME level; treatment cannot be defined separately from the running variable, and attrition from the sample is in this context the outcome of interest. Second, we use the full set of administrative data from DCPS during this period, such that no teacher is omitted from the analysis, regardless of treatment status, and we are therefore able to define retention status for all teachers in the sample, and on both sides of the consequential threshold.

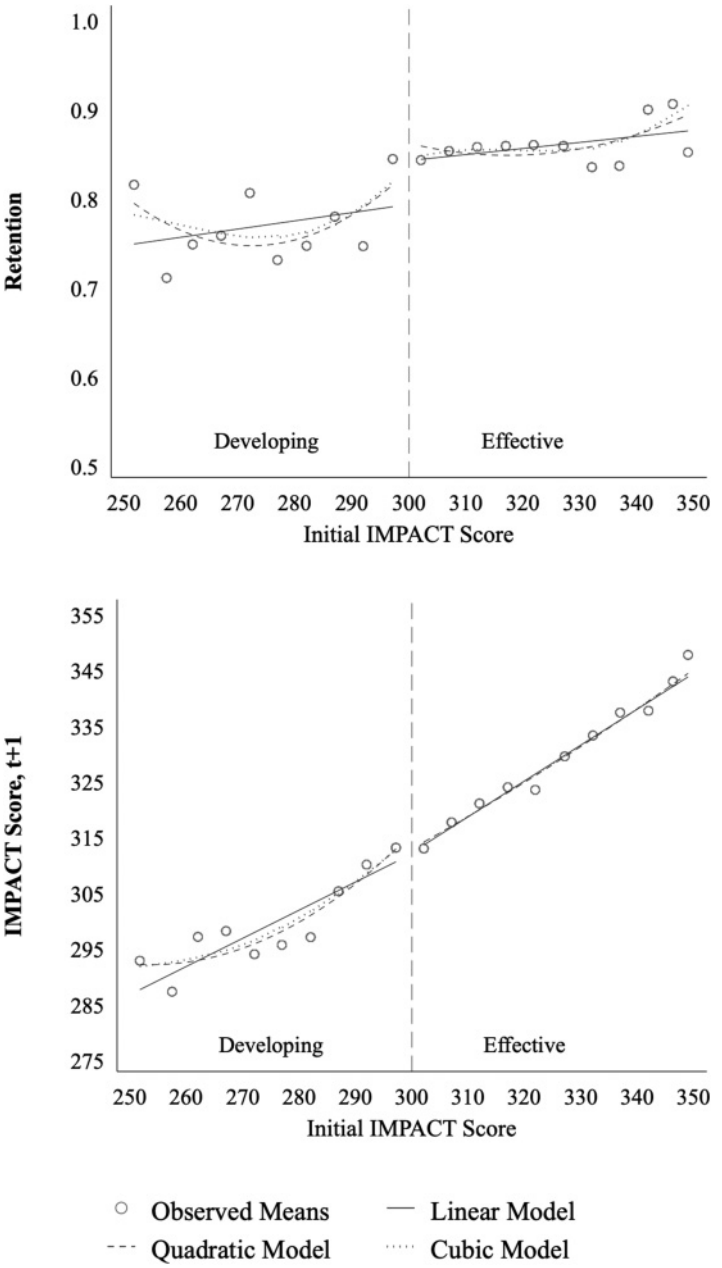
There is risk of differential attrition, however, when examining effects on next-year IMPACT scores. For example, while our administrative data allow us to follow teachers' retention decisions, there are cases in which a teacher might be technically retained in




*Notes:* Each plotted point represents the mean outcome for a given bin (width = 5 IMPACT points) of initial (pre-appeal) IMPACT 3.0 scores. Note that we test for discontinuous retention effects below the D threshold, given there is an apparent drop in the probability of retention for teachers within initial IMPACT scores between 240 and 244. We do this by running a regression with placebo treatment effects at points away from the true cutoff (available in online appendix table A.2), and by testing for differences in mean retention and mean teacher characteristics across bins (not shown); neither test indicates discontinuous effects at any point other than the true threshold.

**Figure 3.** Treatment Effects at the Minimally Effective (ME) Threshold

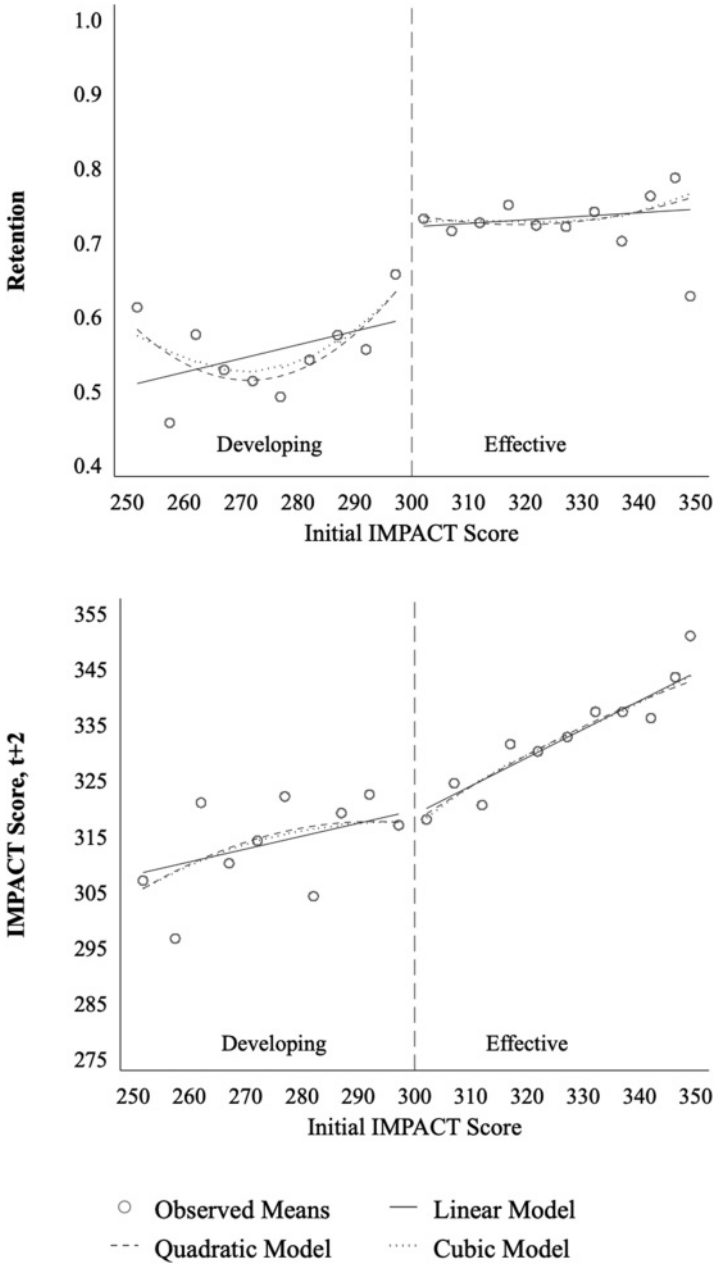
DCPS but not receive IMPACT scores the following year, such as when a teacher goes on maternity leave too early in the academic year to earn an IMPACT score. Our performance estimates would be biased, for example, if there were a differential probability




*Notes:* Each plotted point represents the mean outcome for a given bin (width = 5 IMPACT points) of initial (pre-appeal) IMPACT scores.

**Figure 4.** Treatment Effects at the Developing Threshold, $t + 1$

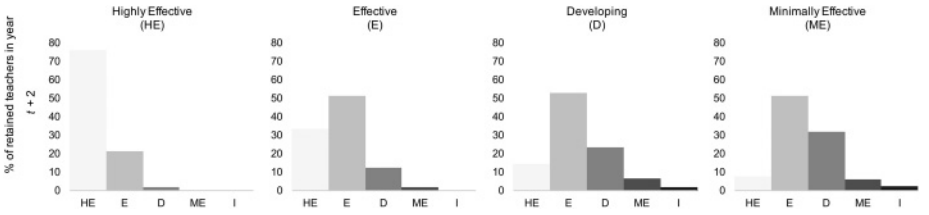
of a teacher not receiving a next-year IMPACT rating across the ME/D threshold, conditional upon being retained in DCPS. We assess this by estimating our analytic model with the probability of receiving a next-year IMPACT score in the left-hand side of the equation. Our estimates indicate that predicted attrition rates for the ME analytic




*Note:* Each plotted point represents the mean outcome for a given bin (width = 5 IMPACT points) of initial (pre-appeal) IMPACT scores.

**Figure 5.** Treatment Effects at the Developing Threshold, $t + 2$

sample are no different (.012, $p = .623$) for treated (.054) and untreated (.042) teachers; across the overall ME analytic sample, 4.42 percent of retained teachers do not receive IMPACT scores the following year. There is similarly no indication of attrition bias within the D analytic sample, where the difference in predicted attrition is less than




*Notes:* Figures exclude teachers rated Ineffective (I), given that an I rating is grounds for immediate dismissal. Fewer than 2 percent of all teachers received an I rating in IMPACT 3.0. Reported ratings are based on teachers' initial IMPACT scores, assigned before the opportunity to appeal for a higher rating. As discussed in the following section, however, few teachers successfully appeal and receive different final scores from those initially assigned.

**Figure 6.** Rating in Year $t + 2$, by Initial Year $t$ Rating

one percentage point (0.001, $p < .834$); treated teachers have an attrition rate of 0.13 percent, compared to 0.12 percent of untreated teachers, or 0.12 percent of the overall D sample.

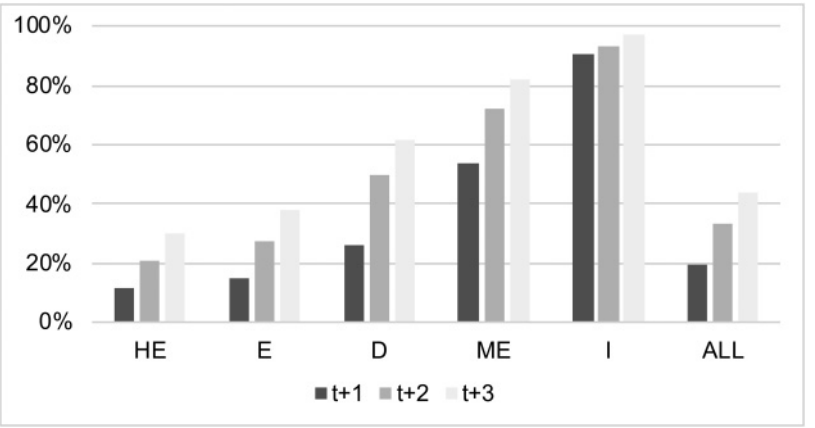
Related to the question of attrition from the analytic sample, one might be concerned that the teachers who are not retained in IMPACT have different improvement potential from those who remain. If, for example, teachers with lower propensity to improve exit at higher rates than their peers, our performance estimates might overstate the gains attributable to IMPACT's performance incentives for all teachers in that rating category. We test for such sorting by estimating, for retained teachers, our RD specification using prior-year performance as the outcome variable. We also compare baseline characteristics across retained versus attrited teachers within a narrow bandwidth of the threshold. We address the results of these tests alongside the corresponding results in the following section.

## 6. RESULTS

### Descriptive Evidence

Most teachers experience meaningful improvement in measured effectiveness over time under IMPACT 3.0. In figure 6, we sort teachers by their initial (pre-appeal) rating in a given year ($t$) and follow their performance over the next two years ($t + 2$). In $t$, most teachers score at least at the Effective level (27.01 percent HE and 43.36 percent E), with about one in five teachers (21.5 percent) scoring at the Developing level, and 6.2 percent achieving a score that places them at the Minimally Effective level. Fewer than 2 percent are rated Ineffective in a given year and these teachers are omitted as they are immediately dismissed. Teachers at each performance level, however, exhibit somewhat different trajectories over the next two years.

Among retained HE teachers, for example, most (76 percent) are still rated HE two years later, and 22 percent are rated E. Few HE teachers (2 percent) receive IMPACT ratings below the E level in year $t + 2$. At the E level, the vast majority of teachers are still earning HE (34 percent) or E (51 percent) ratings two years later, with 12 percent scoring at the Developing level, and 3 percent either ME or I. Developing teachers encompass the new performance category under IMPACT 3.0 that includes a score band under which teachers would have previously been considered Effective. If this category were true to its name, we would expect "developing" teachers to improve their performance




*Notes:* HE = Highly Effective; E = Effective; D = Developing; ME = Minimally Effective; I = Ineffective.

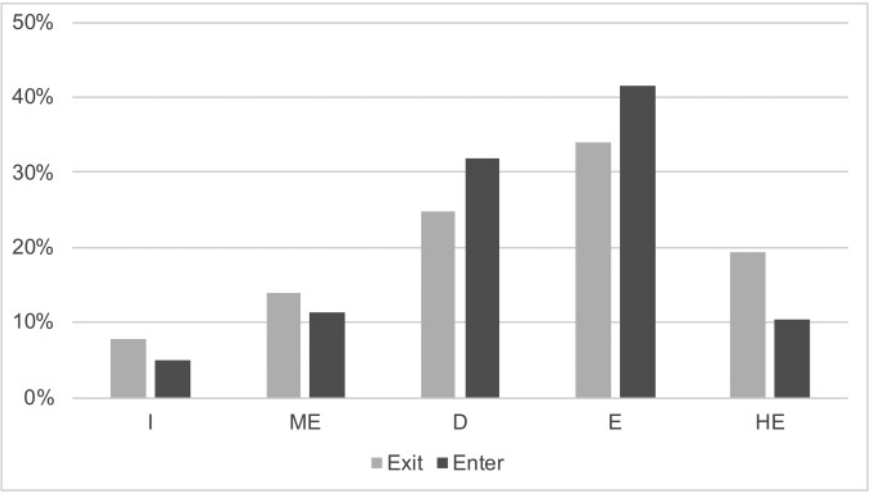
**Figure 7.** Cumulative Attrition of Teachers Over Three Years by IMPACT Rating, 2013–2015

the following year, and indeed this is on average the case for the D teachers who remain. Among the teachers rated D who remained teaching, more than two thirds (68 percent) have improved to E or HE two years later. ME teachers, who make up 6 percent of DCPS educators, not surprisingly—given their incentives—are performing at higher rating levels (91 percent) when they are still teaching in DCPS in year $t + 2$.

Attrition of DCPS teachers is on average relatively high, but exhibits substantial variation depending on IMPACT rating (figure 7). During IMPACT 3.0, nearly 20 percent of all DCPS teachers leave each year, and about 44 percent over three years, which is high compared with some other urban districts (Papay et al. 2017). However, attrition among E and HE teachers is much lower. About 10 percent of HE teachers and 15 percent of E teachers exit DCPS each year with three-year cumulative attrition of 30 and 38 percent, respectively. As might be expected given the incentives of IMPACT, attrition among D, ME, and I teachers is much higher, with one-year attrition of 26, 53, and 91 percent, respectively, and three-year attrition of 62, 82, and 97 percent, respectively. These relatively high levels of attrition may be problematic, especially if DCPS is unable to replace exiting teachers with relatively more effective entering teachers.[6]

On average, DCPS recruits teachers who are roughly comparable to those who exit (figure 8). During IMPACT 2.0, IMPACT involuntarily separated many low-performing teachers and induced substantially more low-performing teachers to voluntarily exit. During this time, the performance of entering teachers exceeded that of exiting teachers (average IMPACT scores for entering teachers were 281 compared with 271 for exiting teachers), as might be expected, because the system exited most of the existing stock of low-performing teachers. Once the stock of low-performing teachers is reduced it is reasonable that DCPS would reflect a pattern more like other urban districts where effectiveness of exiting teachers exceeds that of new teachers (see, for example, Ronfeldt, Loeb, and Wyckoff 2013). Under IMPACT 3.0, the average IMPACT score of

6. For a more detailed examination of teacher turnover in DCPS see James and Wyckoff (2020).




*Notes:* I = Ineffective; ME = Minimally Effective; D = Developing; E = Effective; HE = Highly Effective.

**Figure 8.** Performance Distribution of Exiting, Entering, and All Teachers, IMPACT 3.0

exiting teachers is 296, while that of entering teachers is 294. It is concerning, but not surprising, that the share of HE teachers among exits is nearly twice as prevalent as among entering teachers. Recruiting new teachers who enter as HE is unexpected, as most teachers meaningfully develop over the early years of their careers. It is also not surprising that a smaller percentage of entering teachers are identified as ME (11 percent) or I (5 percent) than among exiting teachers (14 percent ME; 8 percent I) given IMPACT's incentives for very low-performing teachers.

These summaries provide descriptive evidence that: (1) teachers' ratings often improve in IMPACT when they are retained; (2) teachers at lower-performance levels leave at meaningfully higher rates than those with higher IMPACT ratings; and (3) the performance of entering and exiting teachers is roughly comparable in contrast to most urban districts. These descriptive results do not illuminate the extent to which IMPACT *causes* teachers to improve or to voluntarily leave DCPS. The RD analysis that follows explicitly addresses these questions.

### Regression Discontinuity Analysis

#### *First-Stage Effects*

Figure 2 shows the assignment to treatment is not strictly continuous across all IMPACT 3.0 years, due to teachers successfully appealing their IMPACT scores to attain higher ratings. These appeals are concentrated in AY 2012–13, which saw a slightly higher share of successful appeals following an error in the value-added calculation for some teachers, with 6 percent of ME teachers successfully appealing their scores to upgrade to a D rating and 6 percent of D teachers successfully appealing their scores to upgrade to an E or HE rating. For the remaining IMPACT 3.0 years, initial and final rating assignments are nearly strictly discontinuous, with no more than two ME teachers in the ME sample successfully appealing to a higher rating (D) in a given year

and no more than three D teachers in the D sample successfully appealing to a higher rating (E) in a given year.

Regardless, we utilize an intent-to-treat analysis with the assumption—supported by Dee and Wyckoff's (2015) findings—that the threat of dismissal associated with an initial rating of ME would be sufficiently compelling for a teacher to either leave the DCPS teaching force or to stay and improve.

### Retention

#### *Minimally Effective*

Figure 3 provides graphical evidence of large unconditional retention effects (top panel), with far lower average retention among teachers who have scored just below the ME/D threshold in IMPACT 3.0 than those who scored at the D level. When estimated parametrically (table 4), we find these results are large and robust to the inclusion of teacher covariates and school fixed effects, with teachers just below the threshold approximately 11 percentage points less likely to return the following year, an increase in attrition of approximately 40 percent. For reference, these estimates are similar in magnitude to those in IMPACT 2.0, where estimates demonstrate roughly a 9 percentage point decrease in retention (also presented in table 4). These results suggest that IMPACT 3.0 was at least as effective at inducing low-performing teachers to voluntarily exit as it was when initially implemented.

We ran additional analyses to explore the sensitivity of results to varying bandwidths and higher-order polynomials—both tests for the functional form of the relationship between IMPACT scores and retention. The inclusion of a quadratic produces a slightly higher point estimate (14 percent), although the Aikake information criterion (AIC) suggests the linear model with teacher controls and school fixed effects is a slightly better model fit. In addition, we explore the use of triangular-kernel-weighted observations, in lieu of the uniform weights presented in table 4, where greater weight is placed on units closer to the threshold. We find that the use of triangular kernel weights produces estimates at least as large as those with uniform weights (online appendix table A.3), yet our estimates are sensitive to our choice of bandwidth, highlighting the importance of our assumptions about the functional form between teachers' IMPACT scores and retention for estimating internally valid treatment effects. Although larger bandwidths introduce greater precision, they can increase potential bias given that observations farther from the cut point could bias effects seen at the threshold. At the bandwidths that balance squared bias and variance to minimize the asymptotic approximation to the mean-squared error (MSE) of the regression discontinuity point estimator (between 9 and 13 points from the ME/D threshold, depending on the method used; see Cattaneo, Idrobo, and Titiunik 2019), retention effects are even larger—ranging from 21 to 24 percentage points (online appendix table A.3). The estimates at these smaller bandwidths are nearly double that of the estimated retention effect at the full bandwidth (11 percentage points with a bandwidth of $\pm 50$ points). A series of local linear regressions at increasingly smaller bandwidths (online appendix table A.3) show that retention effects are larger at smaller bandwidths, and become smaller as the bandwidth increases to fifty points from the consequential threshold, yet the estimated treatment effects remain substantively large across bandwidth choices, and are significantly different from zero at nearly every bandwidth above a size of ten.

**Table 5.** Reduced-Form Developing Intent-to-Treat Regression Discontinuity Estimates on Teacher Retention and Performance, by Outcome Year

| | Retention | | | | IMPACT Score | | | |
|---|---|---|---|---|---|---|---|---|
| **Outcome Year** | **(1)** | **(2)** | **(3)** | **(4)** | **(1)** | **(2)** | **(3)** | **(4)** |
| **t + 1** | −0.049$^{+}$ | −0.043$^{+}$ | −0.050$^{+}$ | −0.050 | 0.13 | 0.49 | −1.23 | 0.51 |
| | (0.027) | (0.026) | (0.026) | (0.037) | (2.80) | (2.74) | (2.59) | (3.69) |
| | *3,271* | *3,271* | *3,271* | *3,271* | *2,688* | *2,688* | *2,688* | *2,688* |
| Teacher controls | No | Yes | Yes | Yes | No | Yes | Yes | Yes |
| School fixed effects | No | No | Yes | Yes | No | No | Yes | Yes |
| Quadratic of running variable | No | No | No | Yes | No | No | No | Yes |
| AIC | 2,814 | 2,611 | 2,443 | 2,443 | 26,376 | 26,319 | 26,093 | 26,096 |
| **t + 2** | −0.123$^{***}$ | −0.114$^{***}$ | −0.126$^{***}$ | −0.105$^{*}$ | 0.92 | 1.05 | 0.17 | −2.02 |
| | (0.033) | (0.032) | (0.032) | (0.046) | (3.46) | (3.34) | (3.06) | (4.45) |
| | *3,271* | *3,271* | *3,271* | *3,271* | *2,192* | *2,192* | *2,192* | *2,192* |
| Teacher controls | No | Yes | Yes | Yes | No | Yes | Yes | Yes |
| School fixed effects | No | No | Yes | Yes | No | No | Yes | Yes |
| Quadratic of running variable | No | No | No | Yes | No | No | No | Yes |
| AIC | 4,215 | 3,971 | 3,782 | 3,779 | 21,761 | 21,657 | 21,383 | 21,384 |

*Notes:* Robust standard errors are in parentheses and sample sizes are in *italics*. Models include year fixed effects and use uniform kernel weights. Treatment effects are estimated off of teachers who were not rated Developing or Minimally Effective in the prior year. Teacher covariates include gender, race, education, experience, and an indicator for whether the teacher is in a tested grade and subject (Group 1). AIC = Aikake information criterion.

$^{+}p < 0.1$; $^{*}p < 0.05$; $^{***}p < 0.01$.

### *Developing*

Figure 4 provides graphical evidence of small unconditional one-year retention effects (top panel), with somewhat lower average retention among teachers who have scored just below the D/E threshold in IMPACT 3.0 than those who scored at the E level. When estimated parametrically (top panel of table 5), we find these results are robust to the inclusion of teacher covariates and school fixed effects, with teachers just below the threshold approximately 5 percentage points less likely to return the following year, an increase in attrition of approximately 40 percent. Given that teachers rated D have two additional years to earn an E rating or higher, we also follow these teachers' retention into year $t + 2$ (see the top panel in figure 5 and bottom panel of table 5), where the retention effects have compounded relative to E teachers. Across specifications, D teachers are at least 10 percentage points less likely to remain in DCPS by year $t + 2$ than teachers who score at the E level—a similar retention effect to what we observe with teachers just below the ME/D threshold.

We test for the sensitivity of these estimates to functional form assumptions in part by including a quadratic of the running variable (model 4 in table 5), which the AIC suggests is at least as good a model fit as the linear specification (model 3) for both $t + 1$ and $t + 2$. For $t + 1$ retention, the inclusion of the quadratic increases the standard error but does not alter the point estimate (–0.050). For $t + 2$, including a quadratic of the running variable only slightly reduces the estimated retention effect, from −0.126 ($p < 0.001$) to −0.105 ($p < 0.05$). We additionally test for robustness to triangular kernel weights, as well as sensitivity to bandwidth selection. Using both a series of local linear regressions and MSE-minimization estimations, we find that $t + 1$

and $t + 2$ retention estimates are similar in size across most bandwidths. Estimates for $t + 1$ retention change sign but remain small and statistically no different from 0 when the bandwidth is lower than twenty IMPACT points (online appendix table A.4). While $t + 2$ retention effects become statistically insignificant at smaller bandwidths, they remain substantively large across bandwidth and weighting selection.

## Performance

### *Minimally Effective*

The lower panel of figure 3 suggests there may be performance effects from assignment to treatment for ME teachers who choose not to resign from DCPS, with approximately ten points higher average performance among teachers just scoring below D than those just above the threshold. Parametrically, we estimate an IMPACT 3.0 treatment effect of 12.89 IMPACT points in our unconditional model, which becomes an increase of 11.99 points, significant at $\alpha = 0.05$ when we control for teacher covariates and the schools in which they teach. This represents an increase of 27 percent of a SD of IMPACT scores.[7] These performance gains are similar to those observed in the two years of IMPACT 2.0. The inclusion of a quadratic term reduces the size and precision of the estimated performance effect such that it is a no longer statistically distinguishable from zero, though the slightly higher AIC for this model suggests the linear model with teacher controls and school fixed effects is a better fit.

These performance effects are robust to bandwidth choice, with similar estimated treatment effects on next-year IMPACT scores at MSE-optimal bandwidths (between 10 and 11 IMPACT points) to those at the full potential bandwidth (see online appendix table A.3). While performance effects at the ME/D threshold are of similar magnitude across the full range of bandwidths, they are imprecisely estimated even at most larger bandwidths, where the inclusion of additional observations might be expected to improve precision—at best, treatment effects on teachers' next-year performance are significant at $\alpha = 0.10$. Results from these local linear regressions are presented in the top panel of online appendix table A.3. When estimated using triangular kernel weights, effects are also of a similar magnitude (between 7 and 11 IMPACT points), though are statistically insignificant across each model specification.

While we are unable to estimate effects across the subscore components with any precision, particularly for student achievement—which is only available for the limited subset of teachers in tested grades and subjects—analyses available from the authors indicate that teachers at the ME threshold in IMPACT 3.0 make statistically significant gains to the TLF and CSC components of IMPACT. Notably, these are two formative measures where teachers are provided descriptors of exemplary practice, which might make improvements on these two components more feasible than on other measures.

A potential question regarding the generalizability of our overall performance estimates, however, is whether they reflect differential sorting, where ME teachers choose to remain or leave DCPS based on their expected potential to improve. Regardless of

7. The mean IMPACT score for teachers in IMPACT 3.0 is 324, with a standard deviation of 44 IMPACT points.

whether effects are driven by improvements or selection, the policy relevance is the same; a higher-performing teaching force is a key goal of IMPACT, whether achieved through altering the composition or level of teaching in DCPS. Nevertheless, we test for such selection patterns by estimating our RD specification for this sample of teachers, but replacing our outcome variable with lagged IMPACT scores. This test produces small and statistically insignificant effects, suggesting that these performance estimates are not attributable to self-selection. As a secondary test, we limit our respective samples to teachers within a narrower bandwidth from the cutoff (±25 points) and compare baseline characteristics of teachers who remain versus those who leave, to determine whether there might be sorting. We find no difference in terms of previous score gains—which could indicate improvement potential—but small differences in terms of other characteristics (i.e., gender, race, and experience); these are, however, characteristics that we control for in our preferred models.

#### *Developing*

While ME teachers appear to improve their next-year IMPACT scores in response to an immediate dismissal threat, figures 4 and 5 suggest little if any performance effects from assignment to treatment for D teachers who choose not to resign from DCPS. When estimated parametrically, we find null performance effects for both next-year IMPACT scores and the year following (table 5). These null findings persist across model specifications, as well as bandwidth choice and the use of triangular versus uniform kernel weights (online appendix table A.4).

### Other Considerations

It is possible that the overall IMPACT 3.0 intent-to-treat effects we observe on both retention and performance mask heterogeneity in treatment effects by year. We therefore estimated effects on retention and performance by year for each analytic sample (available in tables A.5 and A.6 of the online appendix). Within-year (particularly for retention) ME results are similar in magnitude, though imprecisely estimated. In IMPACT 3.0, our samples decrease substantially due to a combination of compositional changes and the restructuring of rating categories, which shrank the size of our treatment and control score bands. Our by-year estimates of ME treatment effects on teachers' next-year IMPACT scores are fairly stable from year to year, but are in some years more sensitive to decisions about the model specification. Regardless, these by-year estimates, although underpowered, provide suggestive evidence that there may be meaningful ME effects in each year of IMPACT 3.0, and that the overall ME effects we see are not driven by the first year of program revisions. Evidence is a bit more mixed for the D sample, where retention effects in AY 2012–13 and 2014–15 are consistent with across-year results, but anomolous in 2013–14, where there are positive and—in $t + 1$ for all but the quadratic specification—statistically significant effects on retention (approximately 8–9 percentage points) for D teachers relative to E teachers. It is unclear what might have led to different retention effects for teachers receiving their first D rating in 2013–14 relative to other years in IMPACT 3.0, but tests of the equality of retention coefficients across years indicate that these effects are statistically different within the years of our overall analysis. As such, we are cautious about any conclusions regarding IMPACT's retention and performance effects at the D/E threshold.

## 7. DISCUSSION AND CONCLUSION

Ten years ago, reformers touted teacher evaluation as a mechanism to improve teacher effectiveness and student achievement. Despite often-heated debate, virtually every state and school district redesigned its teacher evaluation system in response. Much of the recent public discourse has characterized these reforms as a costly failure that should be abandoned. However, the existing evidence suggests a more nuanced portrait in which these reforms were well implemented and effective in some settings and poorly implemented and ineffective in others. Recent research (Marsh et al. 2017; Donaldson and Woulfin 2018; Cohen et al. 2019) has informed our understanding of this variation in the implementation of teacher evaluation systems (e.g., suggesting the key role of principal take-up). Without a more thorough and rigorous understanding of whether teacher evaluation can improve outcomes for teachers and students across a variety of contexts and how its design and implementation should be altered to be most productive, it seems rash to label it as yet another failed policy.

There is much yet to be learned about the design and implementation of teacher evaluation across a broad set of contexts to realize and sustain its potential. In this paper, we document how the design of IMPACT has changed since its controversial introduction a decade ago and examine whether the initial effectiveness of IMPACT is sustained in the face of major changes in design and context. There are good reasons to believe that these effects may have attenuated in subsequent years. First, the large effects of IMPACT on the improvement in teaching found in AY 2010–11 (Dee and Wyckoff 2015) may have been a singular response to the firings and financial rewards that teachers received in the first year of IMPACT. Second, the context surrounding IMPACT substantially changed over the subsequent eight years. Two new Chancellors and other leadership changes, meaningful design modifications, implementation fatigue and competing priorities, and pressure from stakeholders, all could reduce the effects of IMPACT. The large effects we identify here suggest that rigorous teacher evaluation can be sustained over at least an eight-year period. We observe these effects across years, implying that IMPACT has led to a cumulative improvement in teaching quality and student achievement. These gains benefit students who primarily come from nonwhite, low-income households.

That IMPACT has caused some teachers to improve their skills as measured by TLF is important. The paper shows that IMPACT's differential incentives lead to improved teacher observation (TLF) outcomes. Are such incentives sufficient? Null outcomes from experiments where the treatment is solely teacher pay-for-performance cast doubt on this hypothesis. However, it is more compelling that incentives embedded in a system with strong supports for teacher improvement produce gains in teaching skills. This hypothesis is consistent with our IMPACT findings. Teachers receive multiple classroom observations per year and formal feedback and coaching following each of these evaluations. This feedback may be key to giving teachers the information necessary to make improvements. In fact, analysis of changes in DCPS teaching practice at consequential thresholds under IMPACT 2.0 (Adnot 2016) suggests that teachers strategically improve their practices, as measured by the TLF, when incentivized by IMPACT.

The sustained improvements in teacher effectiveness resulting from IMPACT raise important questions about the national discussion of teacher evaluation. First, an

aspect of improvement in DCPS results from the voluntary exit of teachers who face a dismissal threat. Many districts may find dismissal as employed in DCPS an unrealistic sanction for weak performance. Political or labor market constraints may limit performance-based exits. Evidence from districts confronting different contexts would be very useful.

Second, disillusionment with teacher evaluation reform is largely premised on the observation that there has been little change in the percentage of teachers rated less than effective. We know very little about teachers' behavioral responses to being rated as Effective in a system where there is a Highly Effective category. To what extent do teachers rated as Effective actively engage to improve their performance? Faithfully implementing teacher evaluation is expensive in time and financial resources. Done well, teacher evaluation requires evaluators to be normed and to visit classrooms at least three times during the year. It also requires thoughtful feedback. While evidence on the extent to which states and districts made these investments is limited, it appears doing so may be the exception.

Finally, virtually everyone agrees that differences in teaching effectiveness make a substantial difference for students across a variety of proximal and distal outcomes. Evidence presented in this paper suggests that the rigorous diagnosis of teaching strengths and weaknesses, coupled with feedback intended to improve weaknesses, is a powerful form of professional development. We may disagree about the design of teacher evaluation systems—it is easy to disagree in the face of limited evidence—but it seems difficult to make a persuasive case that teachers should not understand their teaching strengths and weaknesses and be provided with expert feedback on how to improve.

## ACKNOWLEDGMENTS

We are grateful to the District of Columbia Public Schools for supplying the data used in this research and to Scott Thompson, Luke Hostetter, and Lauren Norton for answering our many questions. We appreciate feedback on an earlier version of this paper presented at the Association for Education Finance and Policy and the American Education Research Association meetings. We also benefited from helpful comments from the editors and referees. We received financial support from the Schusterman Family Foundation, the Overdeck Family Foundation, and the Institute of Education Sciences (grants R305H140002 and R305B140026). The views expressed in the paper are solely those of the authors. Any errors are attributable to the authors.

## 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.

Adnot, Melinda. 2016. Teacher evaluation, instructional practice and student achievement: Evidence from the District of Columbia Public Schools and the measures of effective teaching project. PhD dissertation, University of Virginia, Charlottesville, VA.

Adnot, Melinda, Thomas S. Dee, Veronica Katz, and James Wyckoff. 2017. Teacher turnover, teacher quality, and student achievement in DCPS. *Educational Evaluation and Policy Analysis* 39(1): 54–76.

Apperson, Jarod, Carycruz Bueno, and Timothy R. Sass. 2016. Do the cheated ever prosper? The long-run effects of test-score manipulation by teachers on student outcomes. CALDER Working Paper No. 155, American Institutes for Research.

Balch, Ryan, and Matthew G. Springer. 2015. Performance pay, test scores, and student learning objectives. *Economics of Education Review* 44(1): 114–125.

Balingit, Moriah, and Andrew B. Tran. 2018. Before a graduation scandal made headlines, teachers at D.C.'s Ballou High raised an alarm. *Washington Post*, 6 January.

Brown, Emma, Valerie Strauss, and Perry Stein. 2018. It was hailed as the national model for school reform. Then the scandals hit. *The Washington Post*, 10 March.

Cattaneo, Matias D., Nicolás Idrobo, and Rocio Titiunik. 2019. *A practical introduction to regression discontinuity designs: Foundations (Cambridge elements: Quantitative and computational methods for social sciences)*. New York: Cambridge University Press.

Cattaneo, Matias D., Michael Jansson, and Xinwei Ma. 2018. Manipulation testing based on density discontinuity. *Stata Journal* 18(1): 234–261.

Cattaneo, Matias D., Michael Jansson, and Xinwei Ma. 2020. Simple local polynomial density estimators. *Journal of the American Statistical Association* 115(531): 1449–1455.

Chiang, Hanley, Cecilia Speroni, Meriesa Herrmann, Kristin Hallgren, Paul Burkaender, Alison Wellington, and Elizabeth Warner. 2017. *Evaluation of the Teacher Incentive Fund: Final report on implementation and impacts of pay-for-performance across four years*. Available https://ies.ed.gov/ncee/pubs/20184004/pdf/20184004.pdf. Accessed 24 September 2020.

Chuong, Carolyn. 2014. *The inconsistent implementation of teacher evaluation reforms*. Available http://educationnext.org/inconsistent-implementation-teacher-evaluation-reforms/. Accessed 24 September 2020.

Clotfelter, Charles T., Elizabeth Glennie, Helen F. Ladd, and Jacob L. Vigdor. 2008. Would higher salaries keep teachers in high-poverty schools? Evidence from a policy intervention in North Carolina. *Journal of Public Economics* 92(5-6): 1352–1370.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2005. Who teaches whom? Race and the distribution of novice teachers. *Economics of Education Review* 24(4): 377–392.

Cohen, Julie, Susanna Loeb, Luke Miller, and James Wyckoff. 2019. Policy implementation, principal agency, and strategic action: Improving teaching effectiveness in New York City middle schools. *Educational Evaluation and Policy Analysis* 42(1): 134–160.

Cullen, Julie, Cory Koedel, and Eric Parsons. 2021. The compositional effect of rigorous teacher evaluation on workforce quality. *Education Finance and Policy* 16(1): 7–41.

Dee, Thomas, and James Wyckoff. 2015. Incentives, selection, and teacher performance: Evidence from IMPACT. *Journal of Policy Analysis and Management* 34(2): 1–31.

Dee, Thomas S., Will Dobbie, Brian A. Jacob, and Jonah Rockoff. 2019. The causes and consequences of test score manipulation: Evidence from the New York Regents Examinations. *American Economic Journal: Applied Economics* 11(3): 382–423.

Donaldson, Morgaen L., and Sarah Woulfin. 2018. From tinkering to going "rogue": How principals use agency when enacting new teacher evaluation systems. *Educational Evaluation and Policy Analysis* 40(4): 531–556.

Drake, Steven, Amy Auletto, and Joshua M. Cowen. 2019. Grading teachers: Race and gender differences in low evaluation ratings and teacher employment outcomes. *American Educational Research Journal* 56(5): 1800–1833.

Dynarski, Mark. 2016. *Teacher observations have been a waste of time and money*. Washington, DC: Brookings Institution. Available https://www.brookings.edu/research/teacher-observations-have-been-a-waste-of-time-and-money/. Accessed 24 September 2020.

Fixsen, Dean L., Sandra F. Naoom, Karen A. Blase, Robert M. Friedman, and Frances Wallace. 2005. *Implementation research: A synthesis of the literature*. Available https://nirn.fpg.unc.edu/sites/nirn.fpg.unc.edu/files/resources/NIRN-MonographFull-01-2005.pdf. Accessed 24 September.

Fryer, Roland G. 2013. Teacher incentives and student achievement: Evidence from New York City Public Schools. *Journal of Labor Economics* 31(2): 373–407.

Fulbeck, Eleanor S. 2014. Teacher mobility and financial incentives: A descriptive analysis of Denver's ProComp. *Educational Evaluation and Policy Analysis* 36(1): 67–82.

Gates, Bill, and Melinda Gates. 2018. *Annual letter 2018: 10 toughest questions we get asked*. Available https://www.gatesnotes.com/2018-Annual-Letter. Accessed 25 September 2020.

Glazerman, Steven, and Allison Seifullah. 2012. *An evaluation of the Chicago teacher advancement program (Chicago TAP) after four years: Final report*. Washington, DC: Mathematica Policy Research.

Goldhaber, Dan. 2015. Exploring the potential of value-added performance measures to affect the quality of the teacher workforce. *Educational Researcher* 44(2): 87–95.

Gordon, Robert, Thomas J. Kane, and Douglas O. Staiger. 2006. Identifying effective teachers using performance on the job. The Hamilton Project Discussion Paper No. 2006-01. Washington, DC: The Brookings Institution.

Iasevoli, Brenda. 2018. *Teacher-evaluation efforts haven't shown results, say Bill and Melinda Gates*. Available http://blogs.edweek.org/edweek/teacherbeat/2018/02/teacher_evaluation_efforts_haven%27t_shown_results_bill_melinda_gates.html?cmp=soc-edit-tw. Accessed 24 September 2020.

Jacob, Brian A., and Steven D. Levitt. 2003. Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics* 118(1): 843–877.

Jacob, Brian A., Jonah A. Rockoff, Eric S. Taylor, Benjamin Lindy, and Rachel Rosen. 2018. Teacher applicant hiring and teacher performance: Evidence from DC public schools. *Journal of Public Economics* 166:81–97.

James, Jessalynn, and James Wyckoff. 2020. Teacher evaluation and teacher turnover in Equilibrium: Evidence from DC Public Schools. *AERA Open* 6(2): 1–21.

Kane, Thomas J., David Blazar, Hunter Gehlbach, Miriam Greenberg, David Quinn, and Daniel Thal. 2020. Can video technology improve teacher evaluations? An experimental study. *Education Finance and Policy* 15(3): 397–427.

Kraft, Matthew A. 2020. Interpreting effect sizes of education interventions. *Educational Researcher* 49(4): 241–253.

Kraft, Matthew A., and Allison F. Gilmour. 2017. Revisiting *The Widget Effect*: Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Researcher* 46(5): 234–249.

Lee, David S., and Thomas Lemieux. 2009. Regression discontinuity designs in economics. *Journal of Economic Literature* 48(2): 281–355.

Marsh, Julie A., Susan Bush-Mecenas, Katharine O. Strunk, Jane A. Lincove, and Alice Huguet. 2017. Evaluating teachers in the Big Easy: How organizational context shapes policy responses in New Orleans. *Educational Evaluation and Policy Analysis* 39(4): 539–570.

Marsh, Julie A., Matthew G. Springer, Daniel F. McCaffrey, Kun Yuan, Scott Epstein, Julia Koppich, Nidhi Kalra, Catherine DiMartino, and Xiao Peng. 2011. *A big apple for educators: New York City's experiment with schoolwide performance bonuses: Final evaluation report*. Available https://www.rand.org/pubs/monographs/MG1114.html. Accessed 24 September 2020.

McCrary, Justin. 2008. Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics* 142(2): 698–714.

McGee, Kate. 2018. *Most DCPS teachers feel pressure to pass students, teacher union survey says*. Available https://wamu.org/story/18/01/25/dcps-teachers-feel-pressure-pass-students-teacher-union-survey-says/. Accessed 25 September 2020.

McNeil, Michele. 2014. *Race to Top reports detail winners' progress, challenges*. Available https://www.edweek.org/ew/articles/2014/03/19/26rtt.h33.html. Accessed 25 September 2020.

National Assessment of Educational Progress (NAEP). 2007. *The nation's report card: NAEP data explorer*. Available http://nces.ed.gov/nationsreportcard/naepdata. Accessed 30 September 2020.

Papay, John P., Andrew Bacher-Hicks, Lindsay C. Page, and William H. Marinell. 2017. The challenge of teacher retention in urban schools: Evidence of variation from a cross-site analysis. *Educational Researcher* 46(8): 434–448.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. Teachers, schools, and academic achievement. *Econometrica* 73(2): 417–458.

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

Ronfeldt, Matthew, Susanna Loeb, and James Wyckoff. 2013. How teacher turnover harms student achievement. *American Educational Research Journal* 50(1): 4–36.

Rothstein, Jesse. 2015. Teacher quality policy when supply matters. *American Economic Review* 105(1): 100–130.

Ruiz-Primo, Maria A., Richard J. Shavelson, Laura Hamilton, and Steve Klein. 2002. On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching* 39(5): 369–393.

Springer, Matthew G., Dale Ballou, Laura Hamilton, Vi-Nhuan Le, J. R. Lockwood, Daniel F. McCaffrey, Matthew Pepper, and Brian M. Stecher. 2010. *Teacher pay for performance, experimental evidence from the Project on Incentives in Teaching (POINT)*. Nashville, TN: National Center on Performance Incentives, Vanderbilt University.

Springer, Matthew G., John F. Pane, Vi-Nhuan Le, Daniel F. McCaffrey, Susan Freeman Burns, Laura S. Hamilton, and Brian Stecher. 2012. Team pay for performance: Experimental evidence from the Round Rock Pilot Project on team incentives. *Educational Evaluation and Policy Analysis* 34(4): 367–390.

Springer, Matthew G., Walker Swain, and Luis Rodriguez. 2016. Effective teacher retention bonuses: Evidence from Tennessee. *Educational Evaluation and Policy Analysis* 38(2): 199–221.

Stecher, Brian M., Deborah J. Holtzman, Michael S. Garet, Laura S. Hamilton, John Engberg, Elizabeth D. Steiner, Abby Robyn, Matthew D. Baird, Italo A. Gutierrez, Evan D. Peet, Iliana

Brodziak de los Reyes, Kaitlin Fronberg, Gabriel Weinberger, Gerald P. Hunter, and Jay Chambers. 2018. *Improving teaching effectiveness: Final report: The Intensive Partnerships for Effective Teaching through 2015–16*. Available https://www.rand.org/pubs/research_reports/RR2242.html. Accessed 24 September 2020.

Steele, Jennifer L., Richard J. Murnane, and John B. Willett. 2010. Do financial incentives help low-performing schools attract and keep academically talented teachers? Evidence from California. *Journal of Policy Analysis and Management* 29(3): 451–478.

Steinberg, Matthew, and Morgaen Donaldson. 2016. The new educational accountability: Understanding the landscape of teacher evaluation in the post-NCLB era. *Education Finance and Policy* 11(3): 340–359.

Steinberg, Matthew, and Lauren Sartain. 2015. Does teacher evaluation improve school performance? Experimental evidence from Chicago's Excellence in Teaching Project. *Education Finance and Policy* 10(4): 535–572.

Strauss, Valerie. 2015. Teacher evaluation: Going from bad to worse. *The Washington Post*, 1 January.

Taylor, Eric, and John Tyler. 2012. The effect of evaluation on teacher performance. *American Economic Review* 102(7): 3628–3651.

Toch, Thomas. 2018. *A policymaker's playbook: Transforming public school teaching in the nation's capital*. Available https://www.future-ed.org/wp-content/uploads/2018/06/APOLICYMAKERSPLAYBOOK.pdf. Accessed 24 September 2020.

Walsh, Kate, Nithya Joseph, Kelli Lakis, and Sam Lubell. 2017. *Running in place: How new teacher evaluations fail to live up to promises*. Washington, DC: National Council on Teacher Quality.

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*. New York: The New Teacher Project.

What Works Clearinghouse (WWC). 2017. *What Works Clearinghouse standards handbook (version 4.0)*. Available https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf. Accessed 30 September 2020.