Incentives, Selection, and Teacher Performance: Evidence from IMPACT

Thomas S. Dee
James Wyckoff

Abstract

Teachers in the United States are compensated largely on the basis of fixed schedules that reward experience and credentials. However, there is a growing interest in whether performance-based incentives based on rigorous teacher evaluations can improve teacher retention and performance. The evidence available to date has been mixed at best. This study presents novel evidence on this topic based on IMPACT, the controversial teacher-evaluation system introduced in the District of Columbia Public Schools by then-Chancellor Michelle Rhee. IMPACT implemented uniquely high-powered incentives linked to multiple measures of teacher performance (i.e., several structured observational measures as well as test performance). We present regression-discontinuity (RD) estimates that compare the retention and performance outcomes among low-performing teachers whose ratings placed them near the threshold that implied a strong dismissal threat. We also compare outcomes among high-performing teachers whose rating placed them near a threshold that implied an unusually large financial incentive. Our RD results indicate that dismissal threats increased the voluntary attrition of low-performing teachers by 11 percentage points (i.e., more than 50 percent) and improved the performance of teachers who remained by 0.27 of a teacher-level standard deviation. We also find evidence that financial incentives further improved the performance of high-performing teachers (effect size = 0.24). © *2015 by the Association for Public Policy Analysis and Management.*

INTRODUCTION

In recent years, a research consensus has coalesced around the notion that teacher quality is a critically important determinant of student development and achievement (Aaronson, Barrow, & Sander, 2007; Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004; Sanders & Rivers, 1996) as well as later life outcomes (Chetty, Friedman, & Rockoff, 2011). However, there is no similarly wide agreement on how to systematically drive improvements in the quality of the teacher workforce. Districts and schools allocate substantial resources to teacher professional development (e.g., in-service training) despite the fact that there is surprisingly little rigorous evidence on the efficacy of these efforts (e.g., Yoon et al., 2007). Moreover, almost none of this professional development is driven by rigorous assessments of the teaching strengths and weaknesses of individual teachers (Weisberg et al., 2009). Furthermore, decades of empirical research have provided relatively little evidence on observed teacher traits that can consistently predict teacher quality. Nonetheless, the "single-salary" schedules commonly used in U.S. public school districts compensate teachers

Journal of Policy Analysis and Management, Vol. 00, No. 00, 1–31 (2015)

Published by Wiley Periodicals, Inc. View this article online at wileyonlinelibrary.com/journal/pam
DOI:10.1002/pam.21818

according to tightly structured rules that typically reward only teacher experience and education credentials; traits among those without consistent links to teacher quality.

Critics of this status quo argue that such rigid and misaligned compensation systems cannot adequately attract and retain a high-quality teacher workforce (see, e.g., Hanushek, 2007; Johnson & Papay, 2009; Murnane & Olsen, 1989). This misalignment is thought to be especially acute in difficult-to-staff schools where the working conditions are more difficult yet the compensation, due to the single-salary schedule, is often similar to schools with better working conditions. This dissatisfaction has motivated new efforts to design and implement programs to assess and reward teacher performance (Cavanagh, 2011; Johnson & Papay 2009). The enthusiasm for such reforms among some policymakers and some practitioners is underscored by new federal and state initiatives (e.g., the Teacher Incentive Fund, Race to the Top, state waivers from the federal requirements under the No Child Left Behind Act) that promote, among other goals, the design and use of measures of teacher performance in compensation and other personnel decisions. However, these efforts are also extraordinarily controversial and their ongoing implementation appears to be uneven among school districts nationwide. For example, several large urban school districts recently terminated their federally sponsored programs after failing to secure the required buy-in of their teachers' unions (Zubrzycki, 2012). The former New York State Commissioner of Education, John King, recently imposed a teacher assessment system on New York City after the New York City Department of Education and the United Federation of Teachers failed to agree on one, resulting in a loss of $250 million in state aid (Joseph, 2013). More generally, there appears to be renewed resistance to the use of teacher evaluations to assess performance, especially for high-stakes financial and dismissal decisions (McNeil, 2013a; Weiss & Long, 2013).

The heated and ongoing national discussion about reforming teacher evaluation and compensation arguably has its recent genesis in the seminal policy innovations introduced in the District of Columbia Public Schools (DCPS) several years ago under then-Chancellor Michelle Rhee. In the 2009 to 2010 academic year (i.e., AY 2009–10), DCPS introduced IMPACT, a high-stakes teacher evaluation system designed to drive improvements in teacher quality and student achievement. IMPACT established several explicit measures of teacher performance and linked the overall measured performance of individual teachers both to the possibility of large financial incentives as well as to the threat of dismissal. Specifically, during the first three years under this nationally visible program, teachers rated as "highly effective" (HE) have received substantial increases in one time and base compensation, while hundreds of teachers rated as ineffective (or minimally effective [ME] for two consecutive years) have been forcibly separated.

In this study, we utilize unique longitudinal data on DCPS teachers to examine how IMPACT relates to two centrally important policy outcomes: the differential retention of high- and low-performing teachers and subsequent teacher performance conditional on having been retained. In part, we examine this question by presenting descriptive evidence based on the cross-sectional patterns in teacher retention by their measured performance as well as the time series variation in overall teacher performance over the first three years of IMPACT. However, we complement this evidence with inferences based on the strong incentive contrasts embedded within IMPACT.

State and local efforts to provide stronger incentives to teachers are by no means new (e.g., Murnane & Cohen, 1986). A recent body of smaller scale experimental studies (e.g., Springer et al., 2010) suggests that short-term financial incentives linked only to the test performance of a teacher's students are largely ineffective. However, IMPACT has several design features that make it distinctive relative to the

conventional teacher incentives piloted in prior studies. For example, IMPACT has created especially high-powered incentives for teachers; most notably, a *dismissal threat* for low-performing teachers, but also exceptionally large financial rewards for high-performing teachers. This design feature implies that IMPACT targets differential retention of low- and high-performing teachers as well as performance-based financial incentives. A second unique feature of IMPACT is that its incentives are linked to a multidimensional measure of teacher performance (e.g., multiple classroom observations as well as test scores) that is likely to have greater reliability and transparency than test scores alone (e.g., Measures of Effective Teaching [MET], 2013). This targeted performance measure may also enhance the efficacy of IMPACT's incentives because it places some weight on actions teachers readily understand and more directly control (e.g., how their classroom practice relates to defined standards of effective instruction). Third, DCPS provided teachers with support to assist them in meeting IMPACT's expectations (e.g., instructional coaches). Fourth, the incentives created by IMPACT may have stronger credibility for teachers (and better external validity as a policy) because they are part of an at-scale, real-world program that has been sustained over several years rather than a small-scale and temporary experimental pilot.

Unsurprisingly, this dramatic policy innovation in how teachers are evaluated, compensated, and retained is a source of contention that has captured attention nationally. However, there is relatively little empirical evidence on how IMPACT has actually influenced its core proximate outcomes. We present evidence from regression-discontinuity (RD) designs that effectively compare the retention and performance outcomes among teachers whose prior-year performance scores placed them near the threshold values that separated performance ratings (and, by implication, the incentives they faced). For example, teachers whose IMPACT score was 250 to 349 were rated as "Effective" (E) and experienced no unique or immediate consequences with respect to their pay or their job security. In contrast, teachers with scores just below this threshold were rated as Minimally Effective (ME), notified that they would be dismissed if they did not become effective within just one year and did not receive a typical base-pay service credit as indicated on the salary schedule. We present evidence that whether a teacher is just above or below this score threshold can be viewed as conditionally random. This local variation also implies an unusually sharp incentive contrast (i.e., a dismissal threat) that might influence teachers' subsequent retention and performance outcomes.

Another policy-relevant contrast exists among teachers near the 350-point IMPACT score threshold that separates "Effective" from Highly Effective (HE) teachers. Teachers who receive an HE rating immediately qualify for bonus pay. However, they also know that, if they achieve a *second* consecutive HE, they will receive a sizable and permanent increase in their base pay (i.e., equivalent to three to five years of service credit). Such base-pay increases constitute large, durable incentives that are not immediately available to the teachers who scored just below this threshold.

Our RD results indicate that dismissal threats had substantial effects, both increasing the voluntary attrition of low-performing teachers and improving the performance of the previously low-performing teachers who remained within DCPS. Furthermore, our RD design also suggests that financial incentives further improved the performance of high-performing teachers. We assess and discuss both the internal validity threats to these RD designs as well as possible construct-validity concerns related to the performance measures we study. We are also careful to emphasize the stylized nature of the causal estimands that result from these RD designs. In particular, it should be noted that the "localness" of these RD estimates implies that they do not necessarily identify the average treatment effect (ATE) associated with the introduction of IMPACT. However, these results do provide credible evidence on the effects of the types of novel performance incentives

IMPACT introduced. Our study concludes with a discussion of the relevance of this evidence for the ongoing efforts in many states and districts to design and implement new systems of teacher evaluation and compensation.

BACKGROUND

Teacher Evaluation

The practice of teacher assessments has evolved rapidly in recent years. Traditionally, local principals have evaluated the performance of individual teachers using procedures that are fairly superficial, perfunctory, and relatively unstructured. The usual results of such "drive by" assessments are simply to classify individual teachers as either satisfactory or unsatisfactory. These binary designations have typically implied few, if any, direct and meaningful outcomes for teachers (i.e., for compensation, advancement, or professional development). In fact, under these less structured approaches, nearly all teachers are usually rated as satisfactory (Weisberg et al., 2009). However, the policy imperative to more accurately assess the considerable variation in teacher performance has motivated new innovations in the practice of teacher assessment.

The intent of these measures is to accurately and reliably differentiate teacher effectiveness and to provide a basis on which to target a variety of personnel actions (e.g., professional development, tenure, financial rewards, and dismissals). Researchers continue to make progress toward improving the validity and reliability of systems of teacher assessments. Teacher effectiveness in improving student learning is a latent construct, which is related to observable measures, such as teacher value-added and teacher observation rubrics. The research literature has not coalesced around a measure that is agreed to be preferable. However, for teachers to understand how their knowledge and teaching practices can be improved, measures must also be transparent. There is a growing consensus that underscores the possible gains of a balanced approach based on articulating clear and objective standards for teaching practice, relying on multiple sources of data, and employing multiple, carefully trained evaluators (e.g., Danielson & McGreal, 2000; Goe & Croft, 2009; MET, 2013; Pianta & Hamre, 2009). Notably, the final recommendations of the MET project, a three-year study that leveraged a random-assignment design to explore the measurement of effective teaching, provide evidence that teacher effectiveness with respect to test scores can be identified by measures based on past student-achievement gains, rigorous classroom observations, and student surveys (MET, 2013).

The seminal IMPACT teacher-evaluation system, which we describe in more detail below, is broadly consistent with these emerging best practice design principles. However, the evaluation systems currently being implemented in many other school districts appear to remain as works in progress, while public officials continue to grapple with a variety of implementation challenges (e.g., McNeil, 2013b; Ujifusa, 2013). As a result of this ongoing expansion of more rigorous teacher-assessment systems, there is as yet little evidence on their ability to improve teacher performance and student achievement. One exception is Taylor and Tyler (2012) who present evidence, based on the phase-in of teacher evaluations in Cincinnati schools, that merely having a rigorous evaluation (i.e., one with largely informal consequences) improves teacher performance. They find that the students of teachers who have been evaluated improve achievement by 10 percent of a standard deviation more than students of nonevaluated teachers.

Teacher Incentives

Rigid single-salary schedules, which dictate the compensation received by most public school teachers, have been nearly universal in U.S. public schools for well over half of a century. However, throughout this period, there have also been frequent state and local efforts to provide teachers with "merit pay" incentives of various types (Springer, 2009). These initiatives have included teacher rewards for student performance (e.g., test scores or graduation rates), for acquiring skills and certification, and for assuming additional professional responsibilities (i.e., "career ladders"), as well as differentiated compensation for teachers of high-need subjects and in hard-to-staff schools. Proponents of teacher incentives argue that they can drive improvements in student outcomes through multiple channels: (1) by providing financial incentives for teachers to focus or increase their effort, (2) by encouraging the development of stronger teaching skills, (3) by increasing incentives for high-performing teachers to enter or remain in schools subject to the incentives, and (4) by altering the *selection* of individuals into teaching toward those who are more able to benefit from such a reward system.

However, in general, these incentive programs piloted over the last 50 years have been modestly sized and short lived. In a classic article, Murnane and Cohen (1986) argue that the failure of most merit-pay programs for teachers is rooted in a fundamental "evaluation problem." That is, they argued that the support for such initiatives quickly erodes because the inherently "imprecise" nature of effective teaching (e.g., idiosyncratic, multidimensional, and collaborative) renders most types of incentives capricious and demoralizing.[1] In contrast, Ballou (2001) notes that merit pay is used more widely and successfully in private schools, which suggests that there is nothing unique about educational settings that make incentives infeasible. He instead attributes the frequent dismantling of teacher incentives to union opposition.

Despite the prevalence of teacher-compensation reforms, the available empirical evidence on the effects of teacher incentives has, until quite recently, been thin and methodologically weak.[2] However, several recent district-specific studies have provided carefully identified evidence on the extent to which the productivity of existing teachers increases when they are provided with financial incentives (i.e., the first margin enumerated above). For example, the Project on Incentives in Teaching (POINT) was a three-year study that provided randomly assigned middle-school mathematics teachers in Nashville individual bonuses of as much as $15,000 if their students met ambitious performance thresholds (Springer et al., 2010). The availability of these incentives led to no detectable effects on measured student performance or on measures of teacher effort and classroom practice.

A second random-assignment study provided New York City teachers with rewards up to $3,000 for meeting performance targets (Fryer, 2013). In this study, treatment schools had flexibility in designing their incentives and most chose group-based incentives. The impact estimates from this study suggest that the presence of these incentives did not raise school performance and may have even lowered it. A third random-assignment trial of group-based teacher incentives of as much as $6,000 was fielded in a suburban school district in Texas and found no evidence of effects on student outcomes or teachers' attitudes and practices (Springer et al., 2012). A fourth teacher-incentive study set in nine schools outside of Chicago found

[1] However, using data from the Project STAR experiment, Dee and Keys (2004) show that a comparatively sophisticated system (i.e., Tennessee's now-defunct program of financial and career-ladder incentives based on multifaceted evaluations) does generally target rewards to more effective teachers.
[2] For a good overview of this literature, see Springer (2009) or Johnson and Papay (2009).

no effects from conventional individual or group-based incentives of as much as $8,000, but substantial gains in student performance when the incentives were instead framed as a loss rather than a gain (Fryer et al., 2012). Interestingly, the dismissal threats that exist in IMPACT share this "loss aversion" feature.

A fifth study was conducted in 34 Chicago schools that were randomly assigned when (but not if) they implemented the Teacher Advancement Program (TAP). Under this program, teachers were eligible to receive payouts of as much as $6,400 for their contribution to the achievement-based value-added of their students (at the school and school-grade level) and their performance on a classroom observation rubric. Under TAP, teachers could also earn extra pay for undertaking the increased responsibilities associated with promotion to a mentoring or master status. The evidence from this study suggests that random assignment to TAP did not raise student achievement (Glazerman & Seifullah, 2012). However, the program implementation did not occur entirely as intended. Teacher payouts were smaller than the originally stated targets and there were no rewards based on value-added because the requisite linked data systems were inadequate (Glazerman & Seifullah, 2012).

The prevalence of null findings from these recent, district-specific studies obviously raises considerable doubt about the promise of teachers' compensation-based incentives as a lever for driving improvements in teacher performance. One possible explanation for this body of evidence is that teachers already tend to be highly motivated agents for whom additional incentives elicit little behavioral response. Furthermore, it may be that teachers generally lack the willingness (or, possibly, the capacity) to respond to incentives that are linked narrowly and exclusively to test scores.

However, the lack of findings in previous studies may also be driven by design issues. We also note that none of these small-scale experiments have been situated in broad-based strategy for the recruitment, professional development, and retention of effective teachers, especially over the long run. That is, it may be that teacher incentives are more effective when they are viewed as enduring rather than as a temporary pilot. The efficacy of teacher incentives may also turn on the simultaneous presence of professional support and training for teachers. Finally, it could also be that some of the benefits of enduring performance-based compensation for teachers are due to the differential recruitment and retention of high-quality teachers rather than improvements in the performance of extant teachers.

The Structure of IMPACT

In the current context, there are several substantive reasons that IMPACT offers a unique opportunity to examine the effects of a robust package of performance-based teacher incentives. First, as we describe below, IMPACT introduced exceptionally high-powered incentives (i.e., the threat of dismissal for low-performing teachers as well as substantially larger financial incentives for high-performing teachers). Second, these incentives were linked to a multifaceted measure of teacher performance consistent with emerging best practices (e.g., clearly articulated standards, the use of several data sources, including several structured classroom observations), rather than simply to test scores alone. Third, IMPACT also provided teachers with supports (e.g., instructional coaches) to assist them in meeting their prescribed expectations. Fourth, IMPACT is not a small-scale, temporary pilot, but rather a highly visible *at-scale* initiative whose capacity to endure was tested during a contentious mayoral election that coincided with the program's first year.

The basic structure of how teacher performance is measured under IMPACT is relatively straightforward. Following the conclusion of each academic year (i.e., beginning with AY 2009–10), individual DCPS teachers are provided with a single

Table 1. IMPACT score components by teacher type.

Impact component	Teacher type	
	Group 1 (%)	Group 2 (%)
Individual value added (IVA)	50	0
Teaching and learning framework (TLF)	35	75
Teacher-assessed student achievement data (TAS)	0	10
Commitment to the school community (CSC)	10	10
School value added	5	5

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 downwards for "Core Professionalism" (CP) violations reported by principals.

score that summarizes their performance on multiple measures for the academic year (Table 1).

The central component of the overall score for most teachers is based on rigorously scored classroom observations tied to the district's Teaching and Learning Framework (TLF). The TLF specifies the criteria by which DCPS define effective instruction and structures a scoring rubric. The TLF includes multiple domains such as leading well-organized, objective-driven lessons, checking for student understanding, explaining content clearly, and maximizing instructional time.[3] A teacher's TLF score is typically based on five formal observations: three by an administrator (e.g., a principal or assistant principal) and two by a "master educator" (i.e., an expert practitioner who travels across multiple schools to conduct TLF observations independently of administrators). Only the administrator's first observation is announced in advance.

A second component of a teacher's overall score is based exclusively or in part on the test performance of their students. More specifically, for "Group 1" teachers, these scores include their calculated "Individual Value Added" (IVA): a teacher's estimated contribution to the achievement growth of their students as measured on the DC Comprehensive Assessment System (CAS) tests and conditional on student and peer traits.[4] The "Group 1" teachers for whom IVA is calculated are only those for whom the available CAS data allow for the estimation of value added (i.e., only reading and math teachers in grades 4 through 8). The IVA measure is not defined for the majority of DCPS teachers (i.e., about 83 percent of the general-education teachers in DCPS). In lieu of an IVA score, these teachers instead receive a teacher-assessed student-achievement (TAS) score. At the beginning of each academic year, teachers choose (and administrators approve) learning goals based on non-CAS assessments. At the end of the year, administrators rate the teacher's success in meeting these goals using a rubric that emphasizes student learning or content mastery.

[3] In IMPACT's second year, DCPS revised the TLF framework by reducing the number of standards from 13 to 9 and by eliminating some redundancies among these standards. Principal training on the corresponding scoring rubric was also increased.

[4] Teacher value added is converted to a 1 to 4 scale using a conversion table. In 2009–10 and 2010–11, the mean teacher value added was equated to an IVA score of 2.5 with relatively few teachers receiving either a 1.0 or a 4.0. In 2011–12, the mean teacher value-added score was equated to an IVA score of 3.0 and relatively more teachers were assigned to 1 and 4. This had the net effect of increasing average IVA scores by 0.25 in 2011–12. Because of these adjustments, we avoid any year-to-year comparisons for IMPACT scores or their components. Note that this does not affect the within-year comparisons employed in the RD analysis.

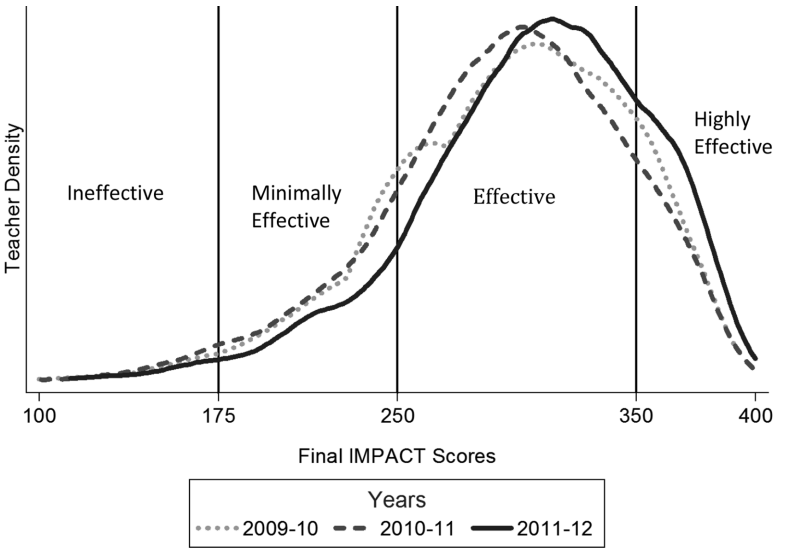


Figure 1. Distribution of IMPACT Scores, AY 2009–10 through AY 2011–12.

All teachers are also assessed by their administrators on a rubric that measures their support of school initiatives, efforts to promote high expectations, and partnerships with students' families and school colleagues: the commitment to the school community (CSC) measure. Teachers also received a score based on their school's estimated value added on the CAS tests (SVA). Finally, principals assess each teacher on their "core professionalism" (CP). The rubric for CP rates teachers on the basis of attendance, punctuality, policies and procedures, and respect. Teachers are assumed to be professionals, and, therefore, CP scores can only reduce a teacher's overall IMPACT score. In AY 2011–12, 12 percent of teachers had their IMPACT scores reduced and these penalties averaged 19 points.

The weighted average of these component scores constitutes a teacher's overall IMPACT score. For the majority of general-education teachers in DCPS (i.e., those in group 2), the TLF observational rubric constitutes 75 percent of their IMPACT score with TAS, CSC, and SVA scores constituting the remainder (Table 1). For the smaller number of group 1 teachers, 50 percent of their overall score is based on their estimated IVA and an additional 25 percent is based on TLF (Table 1). Each component score ranges from 1 to 4 and the overall score is the weighted sum of these, multiplied by 100, so that a teacher's overall score ranges from 100 to 400 prior to possible deductions for CP violations.

These summative IMPACT scores determine high-stakes outcomes for teachers. From AY 2009–10 through AY 2011–12, IMPACT scores allocated teachers to four performance categories: HE teachers (scores of 350 or higher),E teachers (scores from 250 to 349), ME teachers (scores from 175 to 249), and Ineffective (I) teachers (scores below 175). Figure 1 plots the distribution of IMPACT scores for each year.

Those teachers whose score implied an I rating were immediately dismissed. Teachers with an ME rating are subject to a dismissal threat: forcible separation if their next rating is not E or HE. Under "IMPACT*plus*," DCPS also provided rewards to high-performing teachers.

Specifically, from AY 2009–10 through AY 2011–12, IMPACT*plus* provided a one-time bonus to teachers with HE IMPACT ratings. Table 2 shows that these one-time bonuses could amount to as much as $25,000. The size of the bonuses varied based

Table 2. Summary of IMPACT*plus* bonus and base-pay increases.

Bonus pay eligibility	Teachers rated as HE
Bonus pay for teachers in higher poverty schools	$10,000 plus $10,000 for teachers in group 1, plus $5,000 for teachers in a high-need subject
Bonus pay for teachers in lower poverty schools	$5,000 plus $5,000 for teachers in group 1, plus $2,500 for teachers in high-need subject
Base-pay increase eligibility	Teachers rated as HE for a second consecutive year
Base-pay increase for teachers in higher poverty schools	Masters' band + 5-year service credit
Base-pay increase for teachers in lower poverty schools	Masters' band + 3 year service credit

Notes: A higher poverty school is defined as those where the percent of students eligible for free or reduced-price lunches is 60 percent or higher. High-need subjects include special education, bilingual education, and English as a second language as well as secondary math and science. The "Masters' band" implies that the teacher is compensated as if having a master's degree. The exact value of a teacher's base-pay increase following two consecutive HE ratings depends on both their years of experience and their education level. These increases are generally at least $6,000 *per year*. However, salary gain could exceed $20,000 *per year* for teachers without graduate degrees working in high-poverty schools.

on whether the teacher taught in a poor school (defined to be a school where the percentage of free and reduced-price lunch eligible students was at least 60 percent), whether the teacher was in group 1 (teachers with value-added scores), and whether the teacher taught a high-need subject.

Interestingly, IMPACT*plus* also provides strong financial *base-pay* incentives for sustaining high performance. In AY 2009–10 through AY 2011–12, two consecutive years of HE ratings jumped teachers in schools with at least 60 percent free and reduced price lunch eligible students across five years of service credits and the Masters degree lane in the salary schedule. The reward for teachers in schools with fewer than 60 percent of their students eligible for free and reduced price lunch was three years of service and the master's degree lane. The exact magnitude of this base-pay increase also depends on where a teacher is currently situated on the salary schedule. However, these base-pay increases can be as large as $27,000 *per year*. For most teachers, the present discounted value of this permanent pay increase can be substantial. For example, consider a novice teacher just entering employment in DCPS with no prior teaching experience who has a bachelor's degree and currently works in a high-poverty school. At a discount rate of 5 percent (and the differential returns to years of service embedded in the DCPS salary schedule), being twice HE implies salary increases over the next 15 years that are worth $185,259 in current dollars. This is a 29 percent increase in the current value of total earnings over this period. These design features of IMPACT illustrate how the performance bands create sharp incentive contrasts for teachers with scores local to the ME/E threshold (i.e., dismissal threats) and the HE/E threshold (i.e., the possibility of a large base-pay increase). We discuss below the considerable promise of RD designs that can credibly identify the effects of these incentive contrasts on teacher retention and performance.

The effectiveness of the teaching workforce may be improved as a result of compositional changes in teachers and by improving the performance of extant teachers. Design features of IMPACT may affect both teacher composition and teacher performance. The composition of the workforce may lead to greater effectiveness as a result of increased exit of less effective teachers or increased retention of more effective teachers. IMPACT may influence a variety of responses from all existing and prospective teachers. Here, we focus on the effects that result from the differential incentives embedded in IMPACT and which we examine empirically

through RD. Thus, we are concerned with the factors that differentially affect some DCPS teachers and not others.[5] As described above, IMPACT's dismissal threats could cause ME teachers to voluntarily exit at higher rates than would have occurred in the absence of IMPACT; the opportunity for substantial increases in base pay may improve the retention of once HE teachers. Similarly, extant teacher effectiveness may be improved in ME teachers as a result of the dismissal threat or in once HE teachers as a result of the opportunity for base-pay increases.

IMPACT DATA

Our analysis is based on teacher-level administrative data on all DCPS teachers and their students over the first three years of IMPACT (i.e., AY 2009–10 through AY 2011–12). For purposes of comparability, we limit our analytical sample to general-education teachers (i.e., IMPACT groups 1 and 2), who worked in schools that served students in grades K through 12.[6] For each teacher-year observation, we have data on several observed teacher traits, such as race, sex, group status (i.e., IMPACT group 1 or 2), graduate degree, and years of experience (Table 3).[7] We also have several variables characterizing the school in which the teacher worked (e.g., racial ethnic composition, school level, and the share of students eligible for free or reduced-price lunches).

Our data set also contains other teacher-specific data directly related to IMPACT. These include a teacher's IMPACT rating and score as well as their scores on the IMPACT score components (i.e., TLF, IVA, CSC, TAS, and CP). It should be noted that we observe each teacher's initial score and rating as well as their final score and rating, which reflects any repeals or revisions. Such revisions were uncommon, particularly after the first year of IMPACT. Nonetheless, given the potential endogeneity concerns, our RD analysis treats the *initial* IMPACT score and rating as the relevant "intent-to-treat" (ITT) variables (Table 3).

We also used the administrative data available through DCPS to identify whether a teacher rated under the IMPACT system remained employed by DCPS through the next academic year or left for whatever reason (e.g., resignation, retirement, dismissal, or death). This construction means that the two broad outcomes of interest—retention and teacher performance conditional on retention—are observed for two cross-sections of DCPS teachers: AY 2010–11 teacher outcomes as a function of AY 2009–10 IMPACT ratings and AY 2011–12 teacher outcomes as a function of AY 2010–11 IMPACT ratings.

The descriptive evidence we present is based on these annual cross-sections of teachers. That is, in each year, we observe approximately 2,630 teachers.[8] However, several further considerations shaped the samples used in our RD analyses. For

[5] IMPACT may influence the behavior of all DCPS teachers through a variety of mechanisms. For example, all teachers now receive feedback on their performance, which has been shown to improve teacher effectiveness (Taylor & Tyler, 2012). Likewise, the opportunity for a one-time bonus is available to all teachers if they perform sufficiently well in the current year.

[6] This excludes special-education schools and other nonstandard programs as well as teachers with highly specialized assignments (i.e., mostly special-education teachers, but all those serving only English language learners, instructional aides and coaches, teachers of incarcerated youths, etc.).

[7] We constructed teacher experience through cross-referencing repeated cross-sections of several administrative sources (e.g., human-resources data, end-of-year snapshots, and position on the salary schedule). Taken together, these allowed us to develop a more complete and reliable variable.

[8] For purposes of our descriptive evidence, we define teacher retention more finely, distinguishing among teachers who stayed in their school versus transferring as well as whether nonretained teachers left voluntarily (e.g., retirement) or were dismissed. Figure 2 omits teachers who transferred within DCPS to nonteaching positions. In 2009–10 and 2010–11, these teachers constituted 1.7 percent of all teachers in the sample.

Table 3. Descriptive statistics, RD samples.

	ME sample		HE sample	
Variable	Observations	Mean	Observations	Mean
Retained in DCPS, year $t + 1$	4,178	0.84	2,132	0.88
IMPACT score, year $t + 1$	3,447	296.26	1,858	306.70
TLF score, year $t + 1$	3,421	3.03	1,835	3.14
CSC score, year $t + 1$	3,442	3.25	1,855	3.30
TAS score, year $t + 1$	3,349	2.98	1,798	3.10
IVA score, year $t + 1$	632	2.65	300	2.64
CP score, year $t + 1$	3,447	-3.36	1,858	-2.99
ME	4,178	0.16	–	–
ME – ITT	4,178	0.18	–	–
HE	–	–	2,132	0.19
HE – ITT	–	–	2,132	0.19
Initial IMPACT score, year t	4,178	288.62	2,132	314.86
Female teacher	4,178	0.67	2,132	0.68
Teacher sex missing	4,178	0.09	2,132	0.11
Black teacher	4,178	0.52	2,132	0.51
White teacher	4,178	0.28	2,132	0.31
Teacher race missing	4,178	0.14	2,132	0.12
Graduate degree	4,178	0.58	2,132	0.62
Graduate degree missing	4,178	0.12	2,132	0.12
Years of experience: 0–1	4,178	0.21	2,132	0.18
Years of experience: 2–4	4,178	0.16	2,132	0.16
Years of experience: 5–9	4,178	0.18	2,132	0.18
Years of experience: 10–14	4,178	0.12	2,132	0.11
Years of Experience: 15–19	4,178	0.16	2,132	0.18
Group 1 teacher	4,178	0.19	2,132	0.15

Notes: The "ME" sample includes general-education teachers initially assigned an ME or E rating: AY 2009–10 teachers and the AY 2010–11 teachers without a previous ME rating. The "HE" sample includes AY 2009–10 general-education teachers initially assigned to an E or HE rating.

example, for our study of the incentive contrasts that exist at the threshold between ME and effective (E) teachers, we limited the sample to teachers whose initial IMPACT rating placed them in either the ME or E performance bands. This construction allows us to avoid any complications that might be related to other incentive-relevant thresholds in the analytical sample.[9]

An additional complication is that teachers who received a *second* ME rating based on their performance during AY 2010–11 were dismissed automatically under IMPACT. Therefore, their nonretention in DCPS is simply a mechanical effect of this policy rather than voluntary teacher attrition in response to IMPACT incentives. To focus our attention on the choices made by teachers in response to IMPACT's incentives, our RD analysis excludes those AY 2010–11 teachers who had been rated ME in the prior academic year.[10] Overall, this sample construction implies that the RD analysis of the ME threshold is based on 4,178 teacher-by-year observations (Table 3). That is, we observe AY 2010–11 retention and performance outcomes among 2,170 teachers in the ME and E bands during AY 2009–10. And we observe AY

[9] However, including teachers with HE ratings in the analysis of the ME/E threshold leads to similar results, as does including ME teachers in our analysis of the E/HE threshold.

[10] Unsurprisingly, if we instead included the teachers who were forcibly dismissed after a second ME rating, the negative retention effects of an ME rating would appear to be substantially larger.

2011–12 retention and performance outcomes among the 2,008 teachers who were at risk of receiving their first ME rating based on their AY 2010–11 performance.

The analytical sample used in our RD analysis of the threshold that separates effective (E) and HE teachers reflected similar concerns and adjustments. That is, we first limited the sample to teachers whose initial IMPACT rating placed them in the E or HE categories. We also focus exclusively on the first cohort of IMPACT teachers (i.e., AY 2010–11 retention and performance outcomes among the 2,132 teachers rated on their AY 2009–10 performance). Among the subsequent cohort of teachers, an HE rating conflates the mechanical consequences for teachers who had been rated HE in the previous year (i.e., they permanently advance on the salary schedule) with the incentive effects for teachers who received their first HE rating at this time (i.e., they have an opportunity to advance permanently on the salary schedule). Our interest is in the latter effect. However, as it turns out, relatively few teachers ($n = 100$) received their first HE rating based on AY 2010–11 performance (i.e., the large majority of those rated HE had an HE rating in the prior year as well). To avoid obscuring the fact that the identifying variation for the RD analysis of the HE threshold is largely defined for IMPACT's first year, we exclude the second year from our analysis. However, including these data leave our results qualitatively unchanged.

Table 3 presents descriptive statistics for these two analytical samples. We see that the mean teacher retention rate is somewhat lower in the "ME" RD sample (i.e., 84 percent) than in the "HE" RD sample (i.e., 88 percent). Unsurprisingly, the "post-treatment" IMPACT scores are, on average, higher for teachers in the HE analysis than in the ME analysis (i.e., by approximately 10 IMPACT points). However, the other teacher and school-level traits were largely similar across these two samples. Interestingly, the individual value-added (IVA) scores received by teachers were also similar across the ME and HE analytical samples.

As noted earlier, these IVA scores were based on how a teacher's students performed on the DC CAS tests. Allegations of cheating on the DC CAS have received extensive coverage in the press. There are several reasons we believe these allegations are not empirically relevant for the analysis we present here. First and foremost, these test-based measures of teacher performance were only relevant for group 1 teachers under IMPACT and these teachers constitute less than 20 percent of the analytical samples in our RD analysis. Furthermore, our results are robust to excluding these teachers from our analysis. Second, we observe performance separately on all of IMPACT's subcomponents (i.e., IVA and TLF, CSC, TAS, and CP), so we can distinguish performance gains related to CAS scores and those measured in other ways. Third, the most prominent allegations of cheating on the DC CAS actually predate the introduction of IMPACT (Brown, 2013; Gillum & Bellow, 2011). Fourth, during the IMPACT era, DCPS hired independent test-security firms (i.e., Caveon Test Security; Alvarez & Marsal) to assess potential violations. They identified critical violations in no more than a dozen classrooms per year. We have acquired identifiers for the teachers of these classrooms and we find that excluding this quite small number of teachers from our analysis has no practical relevance for the magnitudes or statistical significance of the effects we report.

RD DESIGNS

Our RD analyses effectively compare outcomes among teachers whose *initial* IMPACT scores placed them near the ME/E threshold or near the E/HE threshold. As discussed above, each of these two thresholds implies a sharp and unique contrast in teacher incentives. Teachers who just failed to perform at the effective level face a performance-based employment threat that teachers with effective ratings do not.

Furthermore, teachers who performed just well enough to earn a HE rating have an incentive that effective teachers do not (i.e., the opportunity to earn a permanent increase in base salary).

Our approach to analyzing these discontinuities in teacher incentives has multiple components. Initially, our analysis focuses on basic graphical evidence (Lee & Lemieux, 2009; Schochet et al., 2010). Specifically, we present figures that illustrate how a teacher's final IMPACT rating as well as future outcomes (i.e., retention and performance) vary with the "assignment variable" in this design (i.e., their initial IMPACT score). This graphical evidence provides a compellingly transparent way in which to view this study's key findings as well as some ad hoc guidance relevant to the functional-form considerations for the corresponding regression-based evidence.

We estimate the magnitude and statistical significance of these discontinuities through least-squares specifications that take the following form for outcome Y_i associated with teacher i:

$$Y_i = \alpha I(S_i \leq 0) + f(S_i) + \theta X_i + \varepsilon_i. \quad (1)$$

In this specification, X_i represents teacher covariates and ε_i is a mean-zero random error term. In our preferred specifications, we also condition on fixed effects unique to each of the roughly 120 schools in the analytical samples. The variable, S_i, is the assignment variable (i.e., the teacher's initial IMPACT score) centered on the relevant threshold. Specifically, for our analysis of the effect of ME status on teacher outcomes, we centered teacher's initial IMPACT scores on 249 so that $S_i \leq 0$ implies an "ITT" as an ME teacher. That is, the parameter, α, identifies the "jump" in outcomes for teachers initially rated at or below the ME threshold and conditional on a smooth function of the assignment variable, $f(S_i)$. Our regression-based estimates for the E/HE threshold are similarly structured. However, in those specifications, we centered the initial IMPACT score on 350 and instead estimated the discontinuity that occurs where $I(S_i \geq 0)$. This approach identifies the jump in outcomes for teachers whose initial IMPACT score implied an ITT as an HE teacher.

Our RD analysis also reflects several other considerations and ancillary robustness checks that have been recommended in recent reviews of RD designs (Lee & Lemieux, 2009; Schochet et al., 2010). For example, one key consideration involves the manner in which the regression specification controls for the underlying effects associated with the assignment variable (i.e., $f(S_i)$). In most of the specifications we present, we assume a linear relationship, but allow this to vary above and below the relevant thresholds. Both the graphical evidence and the information criteria from alternative specifications affirm this approach. Nonetheless, we also discuss the results of specifications that condition on higher order polynomials of the assignment variable. Furthermore, our Appendix also presents the results from nonparametric "local linear regressions," which are based on the subset of observations in increasingly tight bandwidths around each threshold as well as nonparametric regressions based on a triangular kernel that places more emphasis on teacher observations nearer to the threshold under study.[11]

The internal validity of all the RD results we present turns on the assumption that whether a teacher was initially assigned above or below a given threshold is conditionally random. One potential threat to this key assumption concerns the possible manipulation of the assignment variable. That is, if some teachers were able to have

[11] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's Web site and use the search engine to locate the article at http://www3.interscience.wiley.com/cgi-bin/jhome/34787.

their initial scores adjusted so that they were systematically able to adjust their initial rating, the RD design outlined here would not provide valid causal inferences. To be clear, the fact that teachers may exert effort to improve their IMPACT scores is not confounding per se (Lee & Lemiuex, 2009). Rather, manipulation would instead invalidate the RD design if teachers with unobserved and outcome-relevant traits were *systematically* able to manipulate their initial rating (i.e., whether their score was above a threshold).

Our institutional knowledge of how initial IMPACT scores were generated (and aggregated) strongly suggests that such manipulation did not occur. However, we also present statistical evidence that speaks to these concerns. For example, in each of our three analytical samples (i.e., HE teachers based on AY 2009–10 performance and ME teachers based on AY 2009–10 and AY 2010–11 performance), density tests (McCrary, 2008) fail to reject the null hypothesis that the distribution of observations is smoothly distributed around each threshold. The absolute values of the test statistics is not larger than 1.17. The panels in Appendix Figure A1 graphically illustrate that teacher observations do not appear to cluster on one side of a threshold (which would have suggested manipulation).[12]

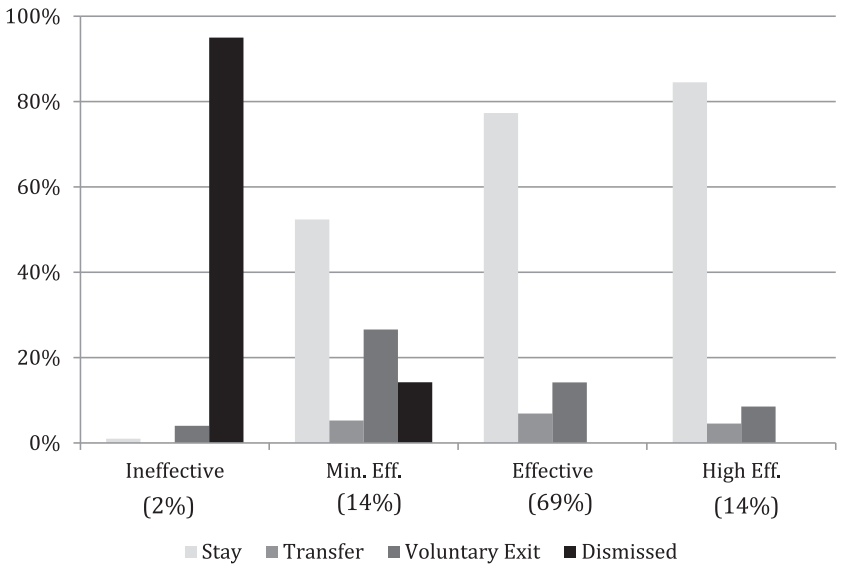
Our Appendix also presents evidence from auxiliary RD regressions that examine the balance of observed teacher traits around each threshold. In the presence of nonrandom sorting around the threshold, we might expect some teacher traits to be clustered on one side of the threshold. However, the regression results in Appendix Table A2 provide evidence that observed teacher traits are quite similar on both sides of these thresholds.[13] That is, in 37 of the 39 regressions for each unique teacher trait and sample, we could not reject the null hypothesis of covariate balance. Two regressions suggested some imbalance of teacher race around the HE threshold. These results could be a multiple-comparison artifact. Regardless, these variables are not significant predictors of teacher performance in these data, implying that they do not constitute a credible internal validity threat. Our Appendix also presents one additional robustness check based on estimating the effects of "placebo" RDs along with the actual threshold relevant under IMPACT.[14] Under the maintained assumptions of the RD design, we would expect the effects of IMPACT's incentive to be concentrated at the 249- and 350-point thresholds that implied a rating change and *not* at other thresholds which have no practical relevance. In our results section, we also discuss potential confounds that are unique to this setting (e.g., nonrandom teacher mobility and rating biases for threatened teachers).

This evidence generally affirms the causal warrant of the RD results we present (i.e., particularly for the effects we find on the ME/E threshold). However, in our final discussion of these RD results, we underscore several important external validity caveats. Arguably, the most important of these concerns the "localness" of the RD estimands. The RD designs used here identify the effects of IMPACT's strong incentive contrasts for the teachers near these thresholds. These local inferences provide an important proof of concept for the role that teacher incentives can play. However, they do not necessarily correspond to an ATE of IMPACT. In contrast, issues related to whether teachers were "compliers" with their original ITT status

[12] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's Web site and use the search engine to locate the article at http://www3.interscience.wiley.com/cgibin/jhome/34787.

[13] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's Web site and use the search engine to locate the article at http://www3.interscience.wiley.com/cgibin/jhome/34787.

[14] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's Web site and use the search engine to locate the article at http://www3.interscience.wiley.com/cgibin/jhome/34787.



Notes: IMPACT ratings are based on performance during AY 2009–10 and AY 2010–11. Retention outcomes are those observed in the subsequent academic years. Units of observation are teacher years and thus teachers may be observed more than once. An "other" retention category, which is always less than 2 percent of any IMPACT rating group, is omitted.

Figure 2. Teacher Retention by IMPACT Rating, AY 2010–11 and AY 2011–12.

under IMPACT have less empirical relevance. For the key effects we report, there is little to no "fuzziness" in the relationship between teachers' initial IMPACT rating and their final ratings (e.g., see Appendix Table A1).[15]

RESULTS

Descriptive Evidence

Relative to typical teacher assessments systems, IMPACT creates substantial differentiation in its teacher ratings. Figure 1 shows the distribution of IMPACT scores for AY 2009–10 through AY 2011–12. In AY 2011–12, 16 percent of teachers earned a HE rating, while 15 percent of teachers are rated I or ME. Between AY 2009–10 and AY 2011—12, mean IMPACT scores improved by 10 points or about 20 percent of a *teacher-level* standard deviation. The improvement in teacher performance is suggestive that IMPACT may have had some of its intended effects. It is also possible that these improvements may have simply resulted from other changes in DCPS that coincided with IMPACT. Figure 2 describes differential retention of teachers during AY 2010–11 and AY 2011–12.

This pattern is also consistent with IMPACT shaping a higher performing workforce. On average, 3.8 percent of all teachers were dismissed as a result of being

[15] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's Web site and use the search engine to locate the article at http://www3.interscience.wiley.com/cgi-bin/jhome/34787.

rated ineffective once or twice ME.[16] In addition to these mechanical dismissals, IMPACT may encourage some low-performing teachers, who otherwise would have remained, to voluntarily exit DCPS. Thirty percent of first-time ME teachers voluntarily exit DCPS, while only 13 percent of teachers who are E or HE do so. As might be expected, ME teachers closest to the effective threshold are more likely to remain in DCPS than those furthest from it. Only 28 percent of first-time ME teachers whose IMPACT scores are within 25 points of the effective threshold (IMPACT scores of 225 to 249) voluntary exit DCPS, while 39 percent of those within 25 points of the ineffective threshold (IMPACT scores of 175 to 199) voluntarily exit. These descriptive outcomes are consistent with a restructuring of the teaching workforce that is implied by the incentives embedded in IMPACT. Less effective teachers under a threat of dismissal are more likely to voluntarily leave than teachers not subject to this threat, and those furthest from the threshold are even more likely to leave. However, other theories of behavior are also consistent with these outcomes. For example, some studies have found that less effective early-career teachers are more likely to exit than more effective novice teachers (Boyd et al., 2011; Goldhaber, Gross, & Player, 2007; Hanushek et al., 2005; Murnane, 1984). We also know from the DCPS data that IMPACT scores for teachers in their first two years of teaching average 17 points less than those with three or more years of experience. Such considerations raise doubts about how to interpret the cross-sectional and time-series evidence from IMPACT. Are we observing the effects of IMPACT incentives or merely observing behavior that would have occurred in the absence of IMPACT? We explore this issue more rigorously employing the RD analysis below.

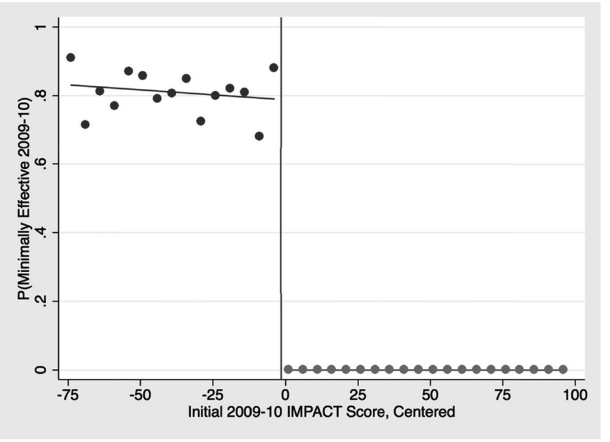
Assignment to Treatment

The logic of a univariate RD design turns in part on the evidence that small changes in an assignment variable lead to large and discontinuous changes in treatment status. With regard to IMPACT, this occurs to the extent that the initial IMPACT scores received by teachers strongly and discontinuously influence their final IMPACT status (and the corresponding incentives they face). In Figure 3, we illustrate these "first-stage" relationships for the discontinuities we study. These figures are based on organizing DCPS teachers into five-point bins based on their initial IMPACT scores (e.g., 245–249, 250–254, etc.) and identifying the share of teachers within these bins with a final status as an ME or HE teacher.

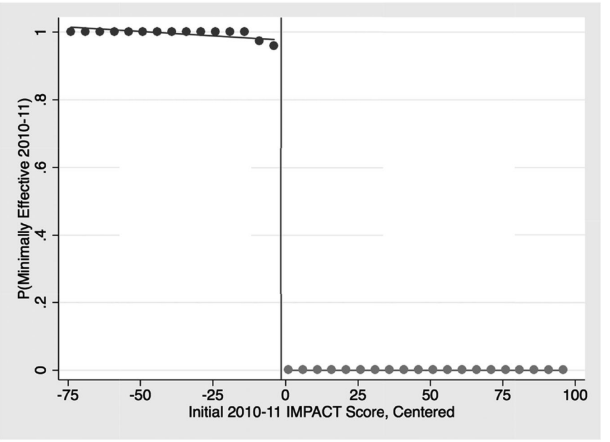
Panel (a) of Figure 3 illustrates this relationship for the first year of IMPACT and ME status. For teachers with initial scores in the effective range (i.e., 250 or higher), the probability of being rated as an ME teacher was zero. However, for teachers with initial IMPACT scores in the ME range, the probability of a final ME rating for AY 2009–10 jumps dramatically to approximately 80 percent. Notably, this relationship reflects some fuzziness: an initial ME rating did not perfectly predict a final ME rating. This is due to the fact that some teachers (i.e., 85 of the 436) were able to appeal successfully their initial IMPACT rating as an ME teacher in IMPACT's first year. Because our research design leverages the variation in incentives generated by *initial* scores, this fuzziness is not an internal validity threat. However, it does suggest the possibility of an external validity caveat: the resulting causal estimands may only be defined for teachers who "complied" with their initial assignment.

This consideration is not relevant for the remaining discontinuities where the relationship between initial scores and teachers' final ratings is "sharp" or virtually so.

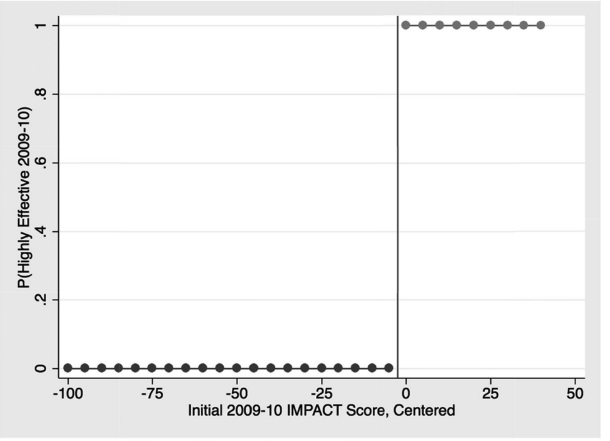
[16] We observe five teachers (0.06 percent of all teachers) rated ineffective who remained due to the appeals process, and eight whose official designation identifies a different form of exit.



(a) Minimally Effective, AY 2009-10



(b) Minimally Effective, AY 2010-11



(c) Highly Effective, AY 2009-10

Notes: Each dot identifies the mean of the variable on the vertical access for teachers whose initial IMPACT score placed them within that five-point bin.

Figure 3. Minimally and HE Assignment, First-Stage.

For example, based on their AY 2010–11 performance, 303 teachers in the analytical sample were initially assigned an ME rating. As panel (b) in Figure 3 indicates, virtually of these teachers (i.e., all but three) retained their ME status after appeals. This contrast across the first two years of IMPACT suggests the District was more flexible in the consideration of appeals of ME status during IMPACT's first year.

However, this flexibility did not extend to HE ratings. Panel (c) in Figure 3 demonstrates that, in IMPACT's first year, there is fully sharp first-stage relationship between initial IMPACT scores and HE status. That is, no teacher in the HE analytical sample changed the IMPACT rating implied by an initial score. In Appendix Table A1, we present the parametric estimates of all the first-stage effects presented in Figure 3.[17] The corresponding standard errors illustrate the precision of these effects and suggest the statistical power of these RD designs to identify reduced-form effects on the outcomes of interest.

Graphical Evidence

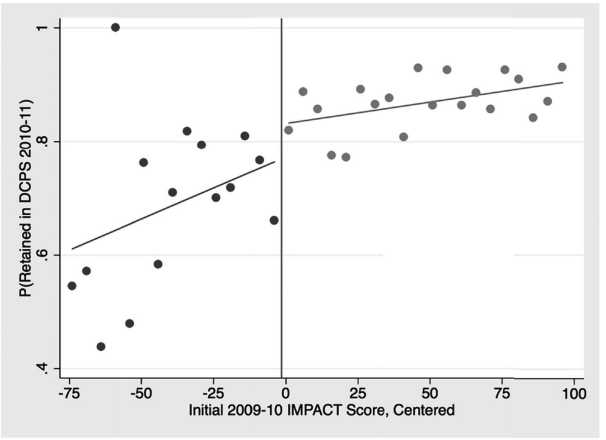
We begin presenting this study's core findings in an unrestrictive and visual manner that closely parallels the first-stage evidence discussed above. That is, Figures 4 and 5 present the conditional means for the next-year teacher outcomes (i.e., retention and performance) as a function of each teacher's *initial* IMPACT score in the prior year. This approach allows us to view how the outcomes of interest vary with the underlying variable that generates strongly discontinuous changes in teacher incentives.

Panel (a) in Figure 4 focuses on teacher retention in AY 2010–11 as a function of their initial AY 2009–10 IMPACT score. This figure illustrates a noticeable drop (i.e., of roughly 5 percentage points) in teacher retention at the threshold that separated ME and effective teachers. This finding suggests that teachers facing a dismissal threat under IMPACT were noticeably more likely to leave voluntarily. The mean retention rate among the teachers in these five-point bins becomes noisier among the lowest performing teachers. However, this reflects in part that there are fewer teachers in the bins that are in the far left of the performance distribution.

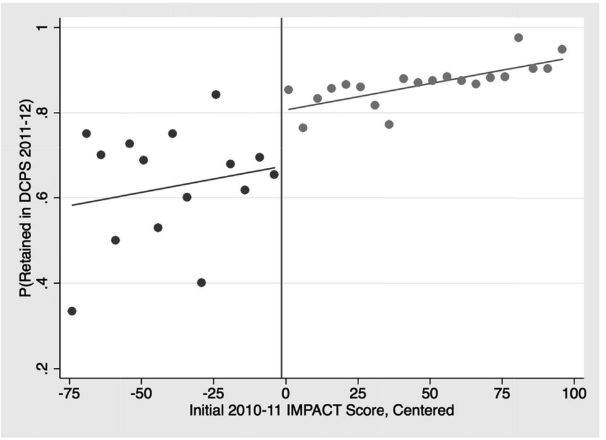
Panel (b) of Figure 4 illustrates the retention effects for teachers near the ME threshold in IMPACT's second year. That is, this figure indicates how the probability a DCPS teacher was retained in AY 2011–12 relates to the initial IMPACT score they received based on their AY 2010–11 performance. It should be noted that teachers were notified of these scores during the summer of 2011. This was the second summer during which teachers who had been rated as ineffective were dismissed and the *first* time that teachers with two consecutive ME ratings were dismissed. Panel (b) indicates that teachers receiving their first ME rating at this time were significantly less likely to return to DCPS for the subsequent academic year. That is, at the threshold where initial IMPACT scores imply an ME rating, we see teacher retention drop by more than 10 percentage points.

Panel (c) of Figure 4 examines the AY 2010–11 retention probabilities for teachers whose initial IMPACT scores for AY 2009–10 placed them proximate to the HE/E threshold. Interestingly, retention during this period was noticeably higher among the higher performing teachers (i.e., near the HE/E threshold, teacher retention was roughly 90 percent). However, this figure suggests that, for teachers just at or above the HE threshold (i.e., those with an opportunity to earn a base-pay increase), retention was higher by approximately 3 percentage points. This pattern is consistent

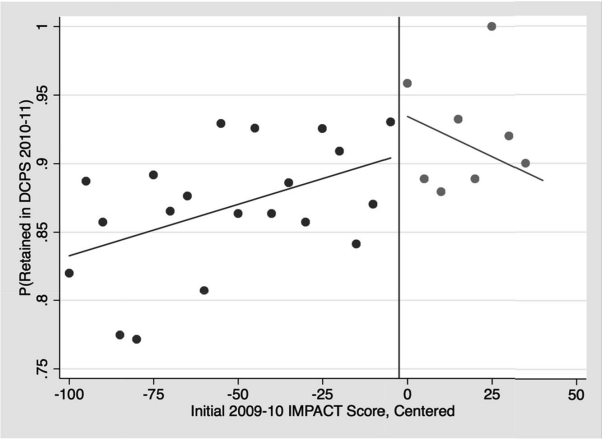
[17] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's Web site and use the search engine to locate the article at http://www3.interscience.wiley.com/cgi-bin/jhome/34787.



(a) Minimally Effective, AY 2009-10



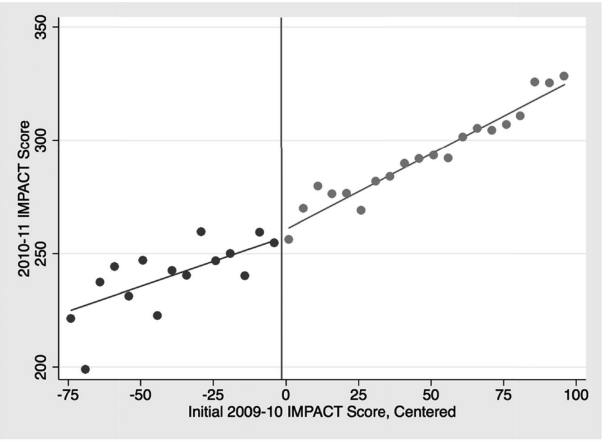
(b) Minimally Effective, AY 2010 to 2011



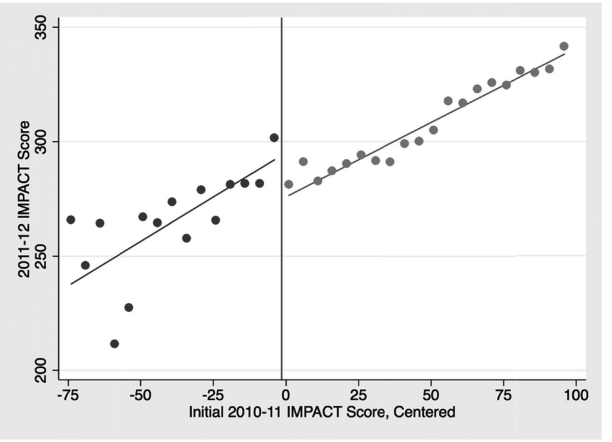
(c) Highly Effective, AY 2009-10

Notes: Each dot identifies the mean of the variable on the vertical access for teachers whose initial IMPACT score placed them within that five-point bin.

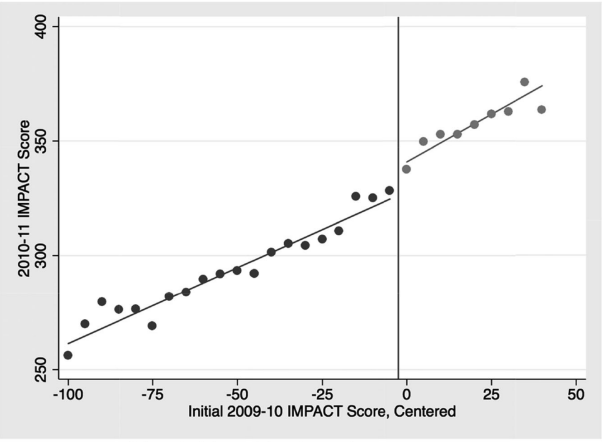
Figure 4. Minimally and HE Retention Effects.



(a) Minimally Effective, AY 2009-10



(b) Minimally Effective, AY 2010-11



(c) Highly Effective, AY 2009-10

Notes: Each dot identifies the mean of the variable on the vertical access for teachers whose initial IMPACT score placed them within that five-point bin.

Figure 5. Minimally and HE Performance Effects.

with the hypothesis that, among higher performing teachers, the opportunity to earn performance-based financial rewards increased retention. In Figure 5, we turn to presenting the *performance* effects of these incentive contrasts for teachers who remained within DCPS.

For example, panel (a) of Figure 5 illustrates how the AY 2010–11 IMPACT scores of teachers relates to their initial AY 2009–10 IMPACT scores. This figure suggests that, in IMPACT's first year, the dismissal threat implied by an ME rating did not induce detectable changes in teacher performance. Panel (b) shows the performance effects of IMPACT's dismissal threats for the second year of IMPACT. That is, panel (b) illustrates how AY 2011–12 teacher performance varied with the incentive contrasts generated by their initial AY 2010–11 performance scores. Notably, these outcomes are measured after the summer of 2011 when DCPS, for the first time, dismissed teachers with consecutive ME ratings.

Interestingly, panel (b) suggests a sizable jump in AY 2011–12 teacher performance (i.e., in excess of 10 points) among those teachers whose initial IMPACT scores placed them under the newly credible dismissal threat implied by an ME rating. This evidence is consistent with the hypothesis that previously low-performing teachers, who remained in DCPS, despite the dismissal threat they faced, undertook steps to meaningfully improve their performance. However, to some extent, the results in panel (b) could indicate that those teachers who had private information about their effectiveness (i.e., that their measured performance would improve even if they behaved no differently) were more likely to stay as DCPS teachers. We suspect that teachers are unlikely to have the sort of information that would allow for this positive selection.[18] Regardless, as a policy matter, this distinction (whether these results reflect teacher improvements or the positive selection of higher quality teachers) is not particularly relevant.

Panel (c) presents evidence on whether AY 2010–11 teacher performance increased for teachers who were initially rated at or above the HE threshold based on their AY 2009–10 performance. These teachers have a powerful financial incentive to continue to perform well because a second consecutive HE rating would imply a permanent increase in base salary. Panel (c) of Figure 5 suggests that there was a noticeable jump in teacher performance (i.e., roughly 10 percentage points) for those who faced these positive financial incentives.

Parametric Results—Retention and Performance

The graphical results discussed above suggest that the dismissal threat implied by an ME rating led to the voluntary attrition of low-performing teachers and improvements in the performance of those who remained (i.e., at least in IMPACT's second year when the dismissal threat implied by ME ratings had established credibility). There is also suggestive evidence that the financial incentives implied by having once been rated HE led to improvements in teacher performance (but not retention). This visual evidence is appealing for several reasons (e.g., its face validity and lack of modeling assumptions). However, it does not allow us to explicitly estimate these effects, to quantify their statistical uncertainty, or to flexibly explore their robustness.

In Table 4, we present the RD estimates that correspond to Figures 4 and 5 and allow for these extensions. The left panel of Table 4 presents the reduced-form RD

[18] An ad hoc empirical decomposition based on our RD design also suggests that incentive effects, rather than selection effects, explain these findings. Using the sample of teachers who returned, we estimated an RD specification where IMPACT performance in the *prior* year is the dependent variable. We find small and statistically insignificant effects that are consistent with the hypothesis of behavioral change in response to incentives.

Table 4. Reduced-form RD estimates, minimally and HE ITT.

	Dependent variable					
	Retained in DCPS, year $t+1$			IMPACT score, year $t+1$		
Sample	(1)	(2)	(3)	(4)	(5)	(6)
	Independent variable: ME ITT					
Full sample	−0.0915***	−0.0675**	−0.0730**	5.841	5.793	4.146
	(0.0318)	(0.0291)	(0.0294)	(3.736)	(3.657)	(3.652)
AY 2009–10	−0.0603	−0.0345	−0.0414	−3.233	−2.200	−2.595
	(0.0423)	(0.0390)	(0.0392)	(5.033)	(4.925)	(4.790)
AY 2010–11	−0.132***	−0.112***	−0.112***	18.35***	16.37***	12.60**
	(0.0481)	(0.0432)	(0.0426)	(5.334)	(5.296)	(5.229)
	Independent variable: HE ITT					
AY 2009–10	0.0263	0.0298	0.0264	12.87***	12.87***	10.93***
	(0.0275)	(0.0236)	(0.0245)	(2.914)	(2.882)	(2.760)
Teacher controls	no	yes	Yes	no	yes	yes
School-fixed effects	no	no	Yes	no	no	yes

Notes: $^{***}P < 0.01$; $^{**}P < 0.05$; $^{*}P < 0.1$. Robust standard errors in parentheses. All models condition on a linear spline of the assignment variable.

estimates where teacher retention is the dependent variable. The first cell in the first row suggests that teachers whose initial IMPACT scores placed them just below the effective threshold were 9 percentage points less likely to be retained. Conditioning on teacher and school-fixed effects reduces this estimate to 7.3 percentage points, but it remains statistically significant. However, the subsequent two rows indicate that these effects were concentrated in the incentives generated by IMPACT's second year.

More specifically, the RD estimates indicate that, in IMPACT's first year (i.e., AY 2009–10), an ME rating reduced teacher retention by a statistically insignificant 3 to 6 percentage points. However, among teachers who received their first ME rating in IMPACT's second year (i.e., AY 2010–11), these retention effects were two to three times larger. That is, an ME rating implied that teacher retention fell by a statistically significant 11 to 13 percentage points. These estimates are quite stable across specifications that introduce teacher controls and school-level fixed effects. One way to frame the magnitude of these effects is to note that just above the ME threshold, roughly 20 percent of teachers did not return to DCPS in the subsequent year. An ME rating that increases this attrition by 11 percentage points implies an increase in teacher attrition of more than 50 percent.

This evidence implies that, in IMPACT's second year (i.e., when the policy was more clearly credible), the dismissal threat implied by an ME rating reduced teacher retention dramatically. Similarly, the bottom left panel suggests that the *positive* financial incentives implied by an HE rating increased teacher retention by roughly 3 percentage points. However, these smaller estimates are not statistically distinguishable from zero.

In the right panel of Table 4, we present the reduced-form RD estimates from specifications where teacher performance as measured by their IMPACT score in the *next* year is the dependent variable. It is worth underscoring here a point made earlier. At least for ME teachers in IMPACT's second year, the incentives created by IMPACT influenced whether a teacher was observed in this analytical sample (i.e., whether they would have an IMPACT score in the year $t+1$). However, in

the presence of this selection effect, these RD estimates have particular relevance because they indicate whether the teachers who chose to remain in DCPS performed at a higher level. The full-sample results in Table 4 suggest that an ME rating had positive, but statistically insignificant effects on IMPACT scores.

However, the subsequent two rows illustrate that these RD results mask the considerable heterogeneity that existed across IMPACT's first two years. An ME rating in IMPACT's first year had small and statistically insignificant effects on subsequent teacher performance. However, in IMPACT's second year, teachers who received ME ratings and chose to remain in DCPS improved their performance in AY 2011–12 by a large and statistically significant amount (i.e., roughly 12.6 IMPACT points in the specification that conditions on school-fixed effects).[19] To put these RD estimates in perspective, it should be noted that the teacher-level standard deviation of AY 2011–12 IMPACT scores among the full sample of groups 1 and 2 teachers is roughly 46. So, these estimates imply an effect size of 0.27 SD (i.e., 12.6/46). The bottom right panel of Table 4 presents estimates based on the HE/E threshold. These estimates similarly indicate that base-pay financial incentives available to teachers on the HE side of the threshold improved subsequent teacher performance by at least 10.9 points (i.e., an effect size of roughly 0.24). Recall that all teachers in DCPS face incentives to improve as all are eligible for substantial one-time bonuses if they are rated HE, and being rated ME leaves open the possibility of subsequent dismissal. As a result, the effects identified above reflect only the differential of incentives for teachers already identified once as either ME or HE.

Because these estimates are based largely on observations of teacher effectiveness at the teacher level, they do not have a conventional interpretation with respect to standard deviations in student-level achievement. However, we can place the magnitudes of these estimates into further perspective in two other ways. One is to note that, for AY 2011–12 teachers who performed near the bottom of the effective range, a gain of 12.6 IMPACT points implies an increase of approximately 5 percentile points (i.e., from the 10th to the 15th percentile) in the distribution of teacher performance. Similarly, for AY 2011–12 teachers at the top of the effective band, a 10.9-point gain is consistent with a 7-percentile increase (i.e., from the 78th to the 85th percentile). A second way to frame these performance gains is to benchmark them against the improvements in performance that are consistently observed during teachers' first three years in the classroom. These gains to experience are typically about 0.07 of a standard deviation of student achievement (Atteberry, Loeb, & Wyckoff, 2013; Clotfelter, Ladd, & Vigdor, 2006; Rivkin, Hanushek, & Kain, 2005). Using a similar approach, we estimate that the typical teacher who entered DCPS in AY 2009–10 with no prior teaching experience improves by 24 IMPACT score points over the first three years of teaching. A gain of 12.6 IMPACT points for teachers at the ME threshold is 52 percent of this three-year gain; the 10.9 gains for teachers at the HE threshold is 41 percent.

Internal and Construct Validity

The RD results presented here suggest that the dismissal threats implied by an ME rating had meaningful effects: inducing voluntary attrition among low-performing

[19] This performance result is robust in specifications that also add quadratic terms for the forcing variable above and below the threshold. The estimated retention effect at the ME threshold also remains large and negative in models that condition on quadratics of the forcing variable though the estimate becomes statistically insignificant because the standard error increases by 40 percent. However, the quadratic measures of the forcing variable are not statistically significant regressors in either model and a specification choice based on information criteria (Schochet et al., 2010) privileges the linear splines.

teachers and improvements in the subsequent performance of those teachers who decided to remain. We also find evidence that, for high-performing teachers, a stronger financial incentive did not induce detectable changes in retention, but did meaningfully improve subsequent teacher performance. Because these RD inferences are identified by small changes in teachers' initial IMPACT scores (in our preferred specifications, among teachers within the *same* schools), they have a credible causal warrant. However, as suggested earlier, we explore the robustness of these causal inferences through several types of evidence that are presented in an Appendix. Density tests (Appendix Figure A1) suggest that these initial scores were not systematically manipulated (i.e., they do not cluster on either side of the threshold). Similarly, teacher covariates are generally balanced around the thresholds (Appendix Table A2). Furthermore, the point estimates associated with the ME threshold are robust as the sample is reduced to increasingly tight bandwidths around that threshold (Appendix Table A3). Furthermore, "placebo" RD estimates indicate that retention and performance effects are *not* found at other thresholds, which did not create incentive contrasts (Appendix Table A5).[20]

The one notable exception to the robustness of these findings concerns the performance effects at the HE/E threshold. In models that limit the sample to tighter bandwidths around this threshold (Appendix Table A4), the magnitude of this effect becomes noticeably smaller, though still sizable and positive. For example, when the sample is limited to the 30 percent of observations within 20 IMPACT points of the HE threshold, the RD estimate falls to 4.3 and becomes statistically insignificant.[21] However, a nonparametric RD regression based on triangular kernel implies that the performance of teachers with an HE rating increases by a weakly significant 5.3 points (Appendix Table A4).[22] The smaller effects associated with tighter bandwidths could reflect the fact that the "control" teachers (i.e., those just below the HE threshold) also experienced quite strong incentives because they had been very close to earning a substantial one-time bonus (as well as the opportunity for a permanent pay increase). Some agnosticism is also suggested because the smaller point estimates also have considerably more statistical uncertainty. Specifically, their 95 percent confidence intervals include the point estimates based on the full sample. Regardless, this finding suggests there is somewhat less certainty about the performance effects at this threshold.

An entirely separate and important set of possible confounds concerns the construct validity of the performance outcomes measured by IMPACT. In particular, there are several theoretically reasonable ways in which the performance effects found here could reflect some type of manipulation or reporting biases rather than true gains in teacher performance. For example, in both RD samples, roughly 8 to 9 percent of the teachers we observe with IMPACT scores in period $t + 1$ earned them in a *different* position (i.e., almost exclusively by teaching in a different school and, in a few cases, through a nonteaching position with IMPACT scores). This teacher mobility could conceivably complicate the performance results presented in Table 4. That is, the teachers facing stronger incentives under IMPACT may have been more likely to seek out different (and possibly more advantageous) assignments, thus inflating their measured performance.

[20] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's Web site and use the search engine to locate the article at http://www3.interscience.wiley.com/cgibin/jhome/34787.

[21] A model that also conditions on a quadratic of the forcing variable suggests a similar decrease in the point estimate and a substantial increase in the standard error.

[22] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's Web site and use the search engine to locate the article at http://www3.interscience.wiley.com/cgi-bin/jhome/34787.

Table 5. RD estimates by IMPACT components, minimally and HE ITT.

	ME			HE
Dependent variable	Full sample	AY 2009–10	AY 2010–11	AY 2009–10
IMPACT score, year $t+1$	4.146	−2.595	12.60**	10.93***
	(3.652)	(4.790)	(5.229)	(2.760)
TLF score, year $t+1$	0.0135	−0.0469	0.0954*	0.117***
	(0.0339)	(0.0451)	(0.0498)	(0.0298)
CSC score, year $t+1$	−0.0176	−0.0361	−0.00114	0.0860***
	(0.0320)	(0.0424)	(0.0461)	(0.0281)
TAS score, year $t+1$	−0.0881	−0.175*	0.0284	0.185***
	(0.0785)	(0.105)	(0.115)	(0.0690)
IVA score, year $t+1$	0.239*	0.0158	0.538**	0.102
	(0.137)	(0.182)	(0.227)	(0.161)
CP score, year $t+1$	0.369	−0.859	1.918*	0.232
	(0.856)	(1.257)	(1.031)	(0.482)

Notes: ***$P < 0.01$; **$P < 0.05$; *$P < 0.1$. Robust standard errors in parentheses. All specifications condition on a linear spline of the assignment variable, the teacher observables, and school-fixed effects (i.e., as in models [3] and [6] in Table 4). TLF, teaching and learning framework; CSC, commitment to school community; TAS, teacher-assessed student learning; IVA, individual value added; CP, core professionalism.

We examined this question directly by estimating auxiliary RD equations in which teacher mobility to a different IMPACT-rated position is the dependent variable. For both the ME and HE thresholds, we could not reject the null hypothesis that IMPACT ratings did not influence teacher mobility. We also find retention and performance effects when the sample is limited to teachers who remain in their original school. An alternative form of possibly confounding teacher mobility would be movement across groups 1 and 2 teaching assignments within their original school. However, auxiliary RD estimates similarly indicate that IMPACT incentives did not have statistically significant effects on teachers' group status in the next year.

Another way in which our core RD results could conceivably be misleading involves whether teachers with strong IMPACT incentives received *biased* reports from their raters. For example, principals are likely to have been aware when one of their teachers faced a dismissal threat due to a prior ME rating or the possibility of a base-salary gain due to a prior HE rating. This awareness may have positively influenced how teachers were rated based on classroom observations (i.e., TLF), on their support for school initiatives (i.e., CSC), on their teacher-assessed student achievement data (i.e., TAS), and on their CP. In Table 5, we present evidence that speaks to these concerns by reporting the RD estimates separately for both the ME and HE thresholds and for each of the IMPACT component scores. The estimates for the full IMPACT scores are also reported again here for reference.

Interestingly, Table 5 indicates that the performance gains observed among teachers with ME ratings from AY 2010–11 are partly due to large improvements in the test performance of students (i.e., the IVA measure). Because raters do not influence these scores, this heterogeneity suggests a limited role for nonrandom reporting bias with respect to the ME results. However, the RD estimates in Table 5 also indicate that teachers facing dismissal threats saw weakly significant improvements in their principal-reported CP (e.g., reduced absenteeism) and in their rated classroom performance (i.e., TLF scores). Interestingly, when we estimate the TLF scores separately for those reported by principals and those reported by external raters (i.e.,

the master educators), the point estimates are almost identical, though less precise. To the extent we believe principals would have a stronger propensity toward reporting biases than district-based raters, this also suggests a limited role for reporting biases. Furthermore, if principals facilitated biased reports for threatened teachers, we might also expect these gains to be observed in higher CSC and TAS scores (but they are not).[23]

The RD estimates in the far right column of Table 5 indicate that the performance gains attributable to HE status were concentrated among TLF, CSC, and TAS scores and not IVA scores. Because each of these affected IMPACT components reflects raters' discretion, the HE results may be more likely to reflect reporting biases. However, at least two observations suggest otherwise. First, if raters were using their discretion to support HE teachers in securing base-pay increases, it is not clear why there were not also statistically significant changes in the CP scores. The absence of effects is not merely due to the lack of CP score penalties in the HE sample. Over 5 percent of the teachers with an initial HE rating in this sample received CP score penalties. Furthermore, the 95 percent confidence interval for this point estimate is sufficiently precise to exclude the gain in CP scores attributable to ME status. Second, RD estimates indicate that HE status led to similarly sized and statistically significant increases in TLF scores when estimated separately by whether the principal or the district-associated master educator was the rater. We would not expect this similarity if reporting biases existed and were stronger among principals than among district-affiliated raters.

CONCLUSIONS AND POLICY IMPLICATIONS

A comparatively strong consensus exists around the notion that teachers have dramatic and long-term effects on the educational and economic outcomes of their students and that there is considerable variance in teacher quality under the current, largely static systems of teacher evaluation and compensation. However, recent studies of teacher-incentive pilots have provided largely discouraging evidence on whether aligning new incentives with singular, test-based measures of teacher performance can improve educational outcomes. This study presents new evidence based on IMPACT, the District of Columbia's controversial teacher evaluation and compensation system that is unique in providing, among other things, exceptionally high-powered, individually targeted incentives linked to performance as measured by multiple sources of data (rather than test scores alone). In this study, we present both descriptive evidence on how IMPACT influences teacher retention and performance as well as RD evidence leveraging the strong incentive contrasts that exist for teachers whose performance placed them near the thresholds for IMPACT's performance bands. Overall, this evidence suggests that IMPACT improved the effectiveness of the DCPS teacher workforce, both through the differential attrition of low-performing teachers and performance gains among those teachers who remained. In particular, the RD estimates provide evidence that the types of incentives that IMPACT created influenced both teacher retention and performance, particularly among lower performing teachers.

Another potentially compelling way to situate these findings more broadly is to contrast them with other carefully identified empirical evidence on alternative policies and practices designed to influence teacher retention and performance. However, we know of relatively few other studies that address this topic with compelling

[23] These null results are not due to ceiling effects in the CSC and TAS ratings. At least 80 percent of the teachers rated as ME in 2011 had CSC and TAS ratings of 3.5 or lower.

research designs. There is some evidence suggesting that practices seeking to promote positive selection into the teaching workforce raise teacher performance. For example, Glazerman, Meyer, and Decker (2006) find that random assignment to a "Teach for America" (TFA) teacher increases student performance by 3 percentile points in math (but has no detectable effects on reading scores). Clotfelter et al. (2008) also find that a bonus for teachers of high-need subjects in high-poverty schools reduced teacher turnover. However, this bonus had no targeting based on teacher performance. There is also some evidence (Glazerman et al., 2010) that a comprehensive induction program providing two years of intensive supports to beginning teachers (e.g., mentoring, classroom observation, and feedback) can improve teacher performance, at least by their third year (but has no detectable effects on teacher retention). A small number of carefully designed studies also suggest that teacher professional development can be effective, though there are far too few to discern patterns in the characteristics of successful programs (Yoon et al., 2007). Clearly, there is much more to be learned about the recruitment, training, development, and retention of higher performing teachers. Nonetheless, in this context, IMPACT appears to be somewhat unique as an initiative that combined multifaceted measurement of teacher performance in the field with high-powered incentives differentially targeting the lowest and highest performing teachers.

Several caveats regarding this study's results are worth underscoring. First and most obviously, because this study's RD estimates leverage the treatment contrasts only for those teachers proximate to performance-band thresholds (and all of whom were subject to IMPACT), they do not necessarily correspond to IMPACT's general effect. Instead, the RD results provide local inferences about the types of incentives that IMPACT created. In addition, we found some evidence that the performance effects for teachers facing dismissal threats were uniquely high for (but not limited to) the smaller number of teachers whose initial scores placed them within just a few points of an effective rating. These threatened teachers are likely to be particularly confident that their subsequent efforts to improve their professional practice would allow them to avoid the consequences of not achieving an effective rating. Interestingly, this treatment heterogeneity dovetails with the conclusions from a larger literature on the design of effective incentive systems in suggesting the critical importance of individuals viewing their targeted tasks as "effort responsive" (e.g., Camerer & Hogarth, 1999). The suggested implications of this for systems of performance-based teacher compensation are worth stressing: the performance of teachers should be more responsive to the incentives they face when they have the knowledge and support to understand how their effort can clearly map into the stated goals. The design of IMPACT appears to reflect these concerns in that the expectations of teachers were clearly articulated and communicated and teaching support to meet these expectations (e.g., instructional coaches) was available.

A notable external validity caveat is that the workforce dynamics under IMPACT may be relatively unique to urban areas, such as District of Columbia, where the effective supply of qualified teachers is comparatively high. A closely related issue is that the contrasts leveraged in this study are among all observed teachers in IMPACT's first three years, which may obscure concerns related to the possible general-equilibrium effects associated with the labor supply of teachers. For example, a simulation study by Rothstein (2012) suggests the teacher firing policies are less effective when they are not accompanied by large salary increases and when performance measurement is noisier. We note that IMPACT coincided with a new teacher contract that provided quite large increases in teacher salaries (Turque, 2010) and that IMPACT also relies on multiple measures of teacher effectiveness, which have been shown to have predictive validity (e.g., MET, 2013). We can also provide some empirical evidence on the dynamics of teacher supply under IMPACT by comparing the performance of teachers who leave and the new hires who replace

them. Teachers who left DCPS at the end of AY 2010–11 had mean IMPACT scores of 255 in their last year, while newly hired teachers for AY 2011–12 averaged 281 in their first year, a difference of about half a standard deviation. However, it may be the case that in other districts there is a smaller supply of potentially effective teachers, constraining the ability of similar policies to improve teacher effectiveness in those districts.

Policymakers are confronted with the difficult decision of how to differentiate among more and less effective teachers. Where to draw these distinctions is even more difficult when there are high stakes associated with the outcomes. Our analysis shows that in a system like IMPACT these differences have important behavioral effects among teachers who are otherwise quite similar. Did DCPS make the right decision in setting the boundaries between ME and E teachers? No one can answer that with certainty. We do note that in AY 2012–13 DCPS divided the previous effective category (IMPACT scores 250–350) that had contained about 70 percent of teachers into a Developing group (IMPACT scores 250–299) and an Effective group (IMPACT scores 300–349), thus raising the performance standard for teachers designated as effective.

A question that is beyond the scope of our analysis is to assess the effects of increased attrition associated with IMPACT on the performance of teachers who remain. While there is evidence that general teacher turnover (transfers and exits) can negatively affect the achievement of students whose teachers remain (Ronfeldt, Loeb, & Wyckoff, 2013), it is unclear whether the exit of relatively ineffective teachers has this effect. For example, it could well be that the performance of other teachers, and their students, improves as less effective colleagues exit and they interact with more effective replacements.

Overall, the evidence presented in this study indicates high-powered incentives linked to multiple indicators of teacher performance can substantially improve the measured performance of the teaching workforce. Nonetheless, implementing such high-stakes teacher-evaluation systems will continue to be fraught with controversy because of the difficult trade-offs they necessarily imply. Any teacher-evaluation system will make some number of objectionable errors in how teachers are rated and in the corresponding consequences they face. Districts may be able to reduce these errors through more sophisticated systems of teacher assessment (e.g., higher-frequency observations with multiple, carefully trained raters), but, in so doing, they will face both implementation challenges and possibly considerable direct financial costs. Policymakers must ultimately weigh these costs against the substantive and long-term educational and economic benefits such systems can create for successive cohorts of students both through avoiding the career-long retention of the lowest performing teachers and through broad increases in teacher performance.

THOMAS S. DEE is a Professor at Stanford University's Graduate School of Education and a Research Associate at the National Bureau of Economic Research, 520 Galvez Mall, CERAS Building, Stanford, CA 94305 (e-mail: tdee@stanford.edu).

JAMES WYCKOFF is the Curry Memorial Professor of Education at the University of Virginia and the Director of the Center on Education Policy and Workforce Competitiveness, 405 Emmet Street South, Charlottesville, VA 22904 (e-mail: wyckoff@virginia.edu).

ACKNOWLEDGMENTS

We received exceptional research assistance from Mindy Adnot and Veronica Katz at the University of Virginia. We are grateful to the D CPS for the data employed in this paper

and to Scott Thompson, Kim Levengood, and Austin Zentz of DCPS for addressing our questions regarding the data and IMPACT. We also received several helpful suggestions from anonymous reviewers. We received financial support for this research from the Carnegie Corporation of New York and the National Center for the Analysis of Longitudinal Data in Education Research (CALDER). CALDER is supported by IES Grant R305A060018. The views expressed in the paper are solely those of the authors and may not reflect those of the funders. Any errors are attributable to the authors.

REFERENCES

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public schools. Journal of Labor Economics, 25, 95–135.

Atteberry, A., Loeb, S., & Wyckoff, J. (2013). Do first impressions matter? Improvement in early career effectiveness. CALDER Working Paper No. 90. Washington, DC: CALDER.

Ballou, D. (2001). Pay for performance in public and private schools. Economics of Education Review, 20, 51–61.

Boyd, D., Lankford, H., Loeb, S., Ronfeldt, M., & Wyckoff, J. (2011). The role of teacher quality in retention and hiring: Using applications-to-transfer to uncover preferences of teachers and schools. Journal of Policy Analysis and Management, 30, 88–110.

Brown, E. (2013). Officials say test cheating in 2008 can't be proved, Washington Post, April 19, 2013, p. B1.

Camerer, C., & Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. Journal of Risk and Uncertainty, 19, 7–42.

Cavanagh, S. (2011). State-by-state battle on bargaining rights continuing to unfold. Education Week, March 9, 2011.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood. Cambridge, MA: National Bureau of Economic Research.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2006). Teacher-student matching and the assessment of teacher effectiveness. Journal of Human Resources, 41, 778–820.

Clotfelter, C., Glennie, E., Ladd, H., & Vigdor, J. (2008). Would higher salaries keep teachers in high-poverty schools? Evidence from a policy intervention in North Carolina. Journal of Public Economics, 92, 1352–1370.

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

Dee, T. S., & Keys, B. (2004). Does merit pay reward good teachers? Evidence from a randomized experiment. Journal of Policy Analysis and Management, 23, 471–488.

Fryer, R. (2013). Teacher incentives and student achievement: Evidence from New York City public schools. Journal of Labor Economics, 31, 373–427.

Fryer, R., Levitt, S., List, J., & Sadoff, S. (2012). Enhancing the efficacy of teacher incentives through loss aversion. NBER Working Paper No. 18237. Cambridge, MA: National Bureau of Economic Research.

Gillum, J., & Bello, M. (2011). When standardized test scores soared in DC, were the gains real? USA Today, March 28, 2011, p. 1A.

Glazerman, S., & Seifullah, A. (2012). An evaluation of the Chicago Teacher Advancement Program (Chicago TAP) after four years. Princeton, NJ: Mathematica Policy Research.

Glazerman, S., Mayer, D., & Decker, P. (2006). Alternative routes to teaching: The impacts of Teach for America on student achievement and other outcomes. Journal of Policy Analysis and Management, 25, 75–96.

Glazerman, S., Isenberg, E., Dolfin, S., Bleeker, M., Johnson, A., Grider, M., & Jacobus, M. (2010). Impacts of comprehensive teacher induction: Final results from a randomized controlled study (NCEE 2010–4028). Washington, DC: National Center for Education

Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Goe, L., & Croft, A. (2009). Methods of evaluating teacher effectiveness. Washington, DC: National Comprehensive Center for Teacher Quality.

Goldhaber, D., Gross, B., & Player, D. (2007). Are public schools really losing their "best"? Assessing the career transitions of teachers and their implications for the quality of the teacher workforce. CALDER Working Paper. Washington, DC: CALDER.

Hanushek, E. (2007). The single salary schedule and other issues of teacher pay. Peabody Journal of Education, 82, 574–586.

Hanushek, E., Kain, J., O'Brien, D., & Rivkin, S. (2005). The market for teacher quality. NBER Working Paper. Cambridge, MA: National Bureau of Economic Research.

Imbens, G. & Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. Review of Economic Studies, 79, 933–959.

Johnson, S. M., & Papay, J. (2009). Redesigning teacher pay. Washington, DC: EPI.

Joseph, C. (2013). New Evaluation System for New York Teachers. New York Times, June 2, 2013, p. A20.

Lee, D. S. & Lemieux, T. (2009). Regression discontinuity designs in economics. Journal of Economic Literature, 48, 281–355.

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

McNeil, M. (2013a). Rifts deepen over direction of Ed. Policy in U.S. Education Week, 32, 1, 14–16.

McNeil, M. (2013b). Feds, states dicker over evaluations. Education Week, 32, 1, 24.

Measures of Effective Teaching (MET) Project (2013). Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET Project's three-year study. Seattle, WA: The Bill and Melinda Gates Foundation.

Murnane, R. (1984). Selection and survival in the teacher labor market. The Review of Economics and Statistics, 66, 513–518.

Murnane, R. J., & Cohen, D. K. (1986). Merit pay and the evaluation problem: Why most merit pay plans fail and a few survive. Harvard Educational Review, 56, 1–17.

Murnane, R. & Olsen, R. (1989). Will there be enough teachers? American Economic Review, 79, 242–246.

Pianta, R. C. & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. Educational Researcher, 38, 109–119.

Rivkin, S., Hanushek, E., & Kain, J. (2005). Teachers, schools, and academic achievement. Econometrica, 73, 417–458.

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

Ronfeldt, M., Loeb, S., & Wyckoff, J. (2013). How teacher turnover harms student achievement. American Educational Research Journal, 50(1), 4–36.

Rothstein, J. (2012). Teacher quality policy when supply matters. NBER Working Paper No. 18419. Cambridge, MA: National Bureau of Economic Research.

Sanders, W. L., & Rivers, J. C. (1996). Research project report: Cumulative and residual effects of teachers on future student academic achievement. Knoxville, TN: University of Tennessee Value-Added Research and Assessment Center,

Schochet, P. Z., Cook, T., Deke, J., Imbens, G., Lockwood, J. R., Porter, J., & Smith, J. (2010). Standards for regression discontinuity designs. Retrieved December 1, 2014, from http://ies.ed.gov/ncee/wwc/pdf/wwc_rd.pdf.

Springer, M. (2009). In M. Springer (Ed.), Rethinking teacher compensation policies: Why now, why again? Performance incentives(chapter 1, pp. 1–21). Washington DC: Brookings Institution.

Springer, M., Ballou, D., Hamilton, L., Le, V., Lockwood, J. R., McCaffrey, D., Pepper, M., & Stecher, B. (2010). Teacher pay for performance, experimental evidence from the project on incentives in teaching. Nashville, TN: National Center on Performance Incentives, Vanderbilt University.

Springer, M., Pane, J., Le, V., McCaffrey, D., Burns, S., Hamilton, L., & Stecher, B. (2012). Team pay for performance: Experimental evidence from the round rock pilot project on team incentives. Educational Evaluation and Policy Analysis, 34, 367–390.

Taylor, E. S. & Tyler, J. H. (2012). The effect of evaluation on teacher performance. American Economic Review, 102, 3628–3651.

Turque, B. D.C. (2010). Teachers likely to ratify contract. Washington Post, June 1, 2010, p. B01.

Ujifusa, A. (2013). Buffalo battle rages over evaluations. Education Week, 32, p. 21.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). The widget effect. Brooklyn, NY: The New Teacher Project.

Weiss, E. & Long, D. (2013). Market-oriented education reforms' rhetoric trumps reality. Broader, bolder approach to education. Retrieved December 1, 2014, from http://www.epi.org/files/2013/bba-rhetoric-trumps-reality.pdf.

Yoon, K. S., Duncan, T., Lee, S. W., Scarloss, B., & Shapley, K. L. (2007). Reviewing the evidence on how teacher professional development affects student achievement: Issues & answers report, REL 2007 No. 033. Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Zubrzycki, J. (2012). Districts abandon grants targeting teacher quality. Education Week, 32, pp. 1, 20–21.

APPENDIX

The three panels in Figure A1 illustrate the densities of teacher observations with respect to RD thresholds in each of the three analytical samples (i.e., HE teachers based on AY 2009–10 performance and ME teachers based on AY 2009–10 and AY 2010–11 performance). Each of these illustrates the lack of a statistically significant difference in the distribution of observations at these thresholds. More specifically, hypothesis tests (McCrary, 2008) confirm that, in each case, we cannot reject the null hypothesis that the discontinuity at these thresholds is zero. The absolute values of these test statistics is not larger than 1.17.

Table A1 presents the first-stage RD estimates for each sample and across different specifications. For the HE inferences, the RD design is "sharp" (i.e., the discontinuity in teachers' initial IMPACT scores perfectly predicts their final status). In the first-year of IMPACT, several successful appeals meant that the relationship between initial IMPACT scores and ME status was somewhat "fuzzy." In IMPACT's second year, the ME first-stage was only modestly fuzzy because there were so few successful appeals of initial IMPACT scores (i.e., only three teachers).

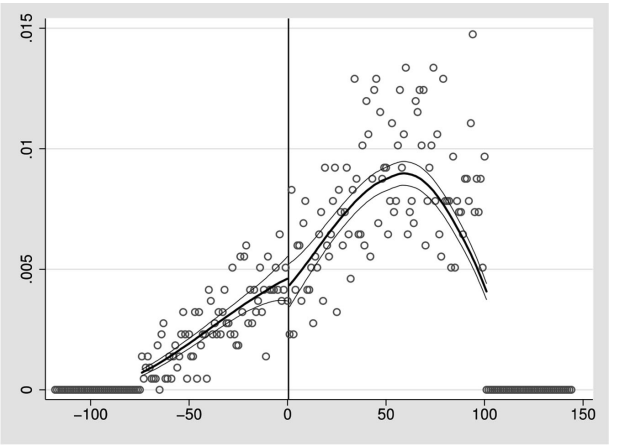
Table A2 presents evidence on the balance of teacher covariates around the RD thresholds. Specifically, this table reports the estimates from 39 individual RD regressions (i.e., 13 for each sample) where a teacher trait is the dependent variable. The prevalence of null results indicates that these teacher traits are similar above and below the RD thresholds that created such strong incentive contrasts for DCPS teachers. The results based on the HE sample indicate a statistically significant imbalance of teacher's race around the HE threshold. However, these results could be viewed as a multiple-comparison artifact (i.e., two statistically significant effects out of 39 inferences). Furthermore, these teacher-race variables are not statistically significant in the IMPACT-score specifications (i.e., Table 4), which implies that this imbalance does not constitute a credible internal validity threat.

Tables A3 and A4 present the core RD results for the ME and HE samples, both for the full-sample and for samples of teachers whose initial IMPACT scores placed them within increasingly tight bandwidths around the relevant thresholds (i.e., 70 points, 60 points, . . . , 20 points). The RD point estimates of the retention effects of these IMPACT incentives remain quite similar as the sample shrinks. However, in the case of the ME results, these point estimates become statistically insignif-

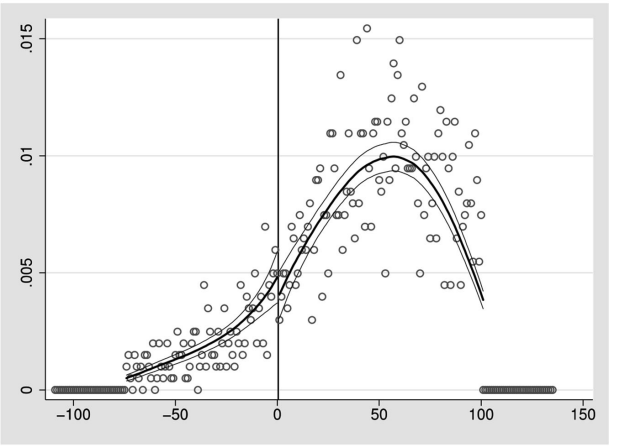
Table A1. First-stage RD estimates, minimally and HE.

Sample	(1)	(2)	(3)
		Dependent variable: ME	
Full sample	0.873***	0.875***	0.874***
	(0.0195)	(0.0194)	(0.0193)
AY 2009–10	0.788***	0.790***	0.790***
	(0.0329)	(0.0324)	(0.0315)
AY 2010–11	0.976***	0.976***	0.976***
	(0.0136)	(0.0136)	(0.0137)
		Dependent variable: HE	
AY 2009–10	1	1	1
	(0)	(0)	(0)
Teacher controls	No	Yes	Yes
School-fixed effects	No	No	Yes

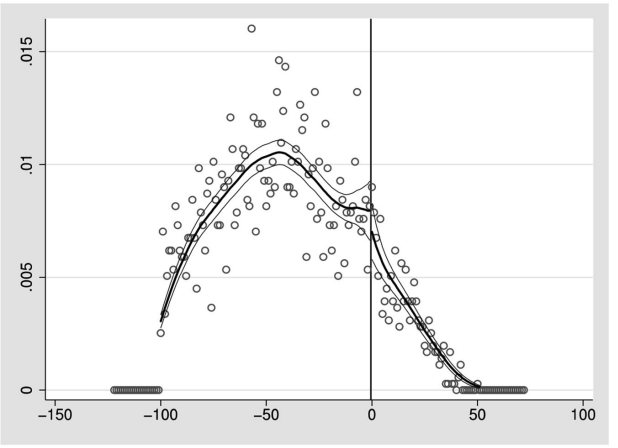
Notes: $^{***}P < 0.01$; $^{**}P < 0.05$; $^{*}P < 0.1$. Robust standard errors in parentheses. All models condition on a linear spline of the assignment variable.



(a) Minimally Effective, AY 2009-10



(b) Minimally Effective, AY 2010-11



(c) Highly Effective, AY 2009-10

Figure A1. Densities of the IMPACT Assignment Variables.

Table A2. Auxiliary RD estimates, covariate balance in the minimally and HE samples.

	ME		HE
Teacher covariate	AY 2009–10	AY 2010–11	AY 2009–10
Female teacher	−0.0295	−0.0345	−0.0177
	(0.0484)	(0.0496)	(0.0428)
Teacher sex missing	0.0552	0.0108	−0.0242
	(0.0407)	(0.0267)	(0.0280)
Black teacher	−0.0411	−0.0415	−0.102**
	(0.0494)	(0.0544)	(0.0451)
White teacher	−0.00524	−0.0345	0.112**
	(0.0408)	(0.0463)	(0.0438)
Teacher race missing	0.0774*	0.0479	−0.0229
	(0.0424)	(0.0433)	(0.0291)
Graduate degree	0.0107	−0.0230	0.0127
	(0.0505)	(0.0559)	(0.0468)
Graduate degree missing	0.0568	0.0347	−0.00395
	(0.0409)	(0.0370)	(0.0303)
Years of experience: zero to one	−0.0771*	0.0243	−0.0140
	(0.0421)	(0.0489)	(0.0307)
Years of experience: two to four	0.0170	−0.0402	−0.00309
	(0.0332)	(0.0399)	(0.0362)
Years of experience: five to nine	0.0155	0.0710*	0.0163
	(0.0388)	(0.0404)	(0.0427)
Years of experience: 10 to 14	0.0112	0.0338	0.00867
	(0.0298)	(0.0369)	(0.0330)
Years of experience: 15 to 19	−0.0127	−0.0326	−0.0429
	(0.0365)	(0.0403)	(0.0376)
Group 1 teacher	−0.0670*	0.00478	0.00229
	(0.0402)	(0.0461)	(0.0307)

Notes: $^{***}P < 0.01$; $^{**}P < 0.05$; $^{*}P < 0.1$. Robust standard errors in parentheses. All models condition on a linear spline of the assignment variable and school-fixed effects.

icant as the shrinking sample sizes imply larger standard errors. The estimated performance effects of an ME threat shrink somewhat (and become statistically insignificant) when we limit the observations to teachers whose initial scores placed them within 40 or 50 points of the ME threshold. However, these estimated effects become somewhat larger and have weak statistical significance when the sample is further limited to teachers within 20 or 30 points of the ME/E threshold and in nonparametric regressions based on a triangular kernel.

A procedure recently developed by Imbens and Kalyaranaman (2012) constructs RD estimates using a bandwidth chosen to balance the precision loss and unbiasedness gains of smaller samples. When applied to these data, the IK procedure suggests a quite narrow bandwidth based only on the $n = 122$ observations within nine points of the ME threshold (only 56 of these teachers had initial scores in the ME range). The estimated performance gain implied by this narrow bandwidth is quite large (46.9 points) and statistically significant. Graphically, the unique performance gains for the small set of teachers close to the ME threshold can be seen in panel (b) of Figure 5 (i.e., the five-point bin just to the left of the threshold). In the conclusion, we note this treatment heterogeneity and its implications for program design in this context. However, we should also note that, when these 122 observations are excluded (i.e., a "donut hole" RD approach), an ME threat still implies a 10-point performance gain.

Table A3. Reduced-form RD estimates, 2010–11 ME ITT, by alternative bandwidths.

Bandwidth	Dependent variable: Retained in DCPS, AY 2011–12		Dependent variable: 2011–12 IMPACT score	
Sample	*n*	Estimate	*n*	Estimate
Full sample	2,008	−0.112***	1,647	12.60**
		(0.0426)		(5.229)
$\lvert S_i \rvert \leq 70$	1,493	−0.109**	1,186	11.83**
		(0.0463)		(5.626)
$\lvert S_i \rvert \leq 60$	1,278	−0.0952*	1,008	12.90**
		(0.0502)		(6.039)
$\lvert S_i \rvert \leq 50$	1,043	−0.0852	812	7.134
		(0.0536)		(6.562)
$\lvert S_i \rvert \leq 40$	804	−0.0855	617	10.52
		(0.0603)		(7.157)
$\lvert S_i \rvert \leq 30$	580	−0.0274	445	15.57*
		(0.0692)		(8.795)
$\lvert S_i \rvert \leq 20$	384	−0.136	289	21.50*
		(0.0942)		(12.59)
Kernel regression	783	−0.0610	602	13.14*
		(0.0606)		(7.153)

Notes: $^{***}P < 0.01$; $^{**}P < 0.05$; $^{*}P < 0.1$. Robust standard errors in parentheses. All models condition on a linear spline of the assignment variable, the teacher observables, and school-fixed effects. The kernel regressions are based on a triangular kernel weight.

Table A4. Reduced-form RD estimates, 2009–10 HE ITT, by alternative bandwidths.

Bandwidth	Dependent variable: Retained in DCPS, AY 2010–11		Dependent variable: 2010–11 IMPACT score	
Sample	*n*	Estimate	*n*	Estimate
Full sample	2,132	0.0264	1,858	10.93***
		(0.0245)		(2.760)
$\lvert S_i \rvert \leq 70$	1,747	0.0365	1,542	8.585***
		(0.0255)		(2.903)
$\lvert S_i \rvert \leq 60$	1,569	0.0379	1,389	7.796***
		(0.0261)		(2.982)
$\lvert S_i \rvert \leq 50$	1,361	0.0503*	1,211	6.255**
		(0.0273)		(3.003)
$\lvert S_i \rvert \leq 40$	1,148	0.0281	1,022	7.501**
		(0.0288)		(3.052)
$\lvert S_i \rvert \leq 30$	915	0.0236	816	6.183*
		(0.0320)		(3.434)
$\lvert S_i \rvert \leq 20$	634	0.0313	566	4.278
		(0.0407)		(3.943)
Kernel regression	1,025	0.0342	913	5.371*
		(0.0297)		(3.154)

Notes: $^{***}P < 0.01$; $^{**}P < 0.05$; $^{*}P < 0.1$. Robust standard errors in parentheses. All models condition on a linear spline of the assignment variable, the teacher observables, and school-fixed effects.

Table A5. Placebo RD estimates.

Independent	ME, AY 2010–11		HE, AY 2009–10	
Variable	Retained	IMPACT score	Retained	IMPACT score
$S_i \leq -20$	−0.0798	−13.81	0.0686	−3.597
	(0.103)	(13.30)	(0.0453)	(4.864)
$S_i \leq -15$	0.126	−3.545	−0.0141	−4.252
	(0.117)	(13.34)	(0.0546)	(5.435)
$S_i \leq -10$	−0.137	3.987	−0.0235	2.830
	(0.109)	(11.11)	(0.0421)	(4.126)
$S_i \leq -5$	0.0317	−11.22	−0.00982	−3.983
	(0.0880)	(11.05)	(0.0435)	(4.382)
$S_i \leq 0$ (actual RD)	−0.154**	16.99*	0.0369	6.652
	(0.0728)	(9.918)	(0.0403)	(4.387)
$S_i \leq 5$	0.106	−8.629	0.0498	2.126
	(0.0701)	(8.875)	(0.0493)	(5.084)
$S_i \leq 10$	−0.0295	9.491	0.0173	−4.381
	(0.0675)	(8.543)	(0.0622)	(5.215)
$S_i \leq 15$	−0.00734	1.581	−0.0409	1.236
	(0.0618)	(7.985)	(0.0577)	(5.112)
$S_i \leq 20$	−0.0372	2.143	0.0186	−0.975
	(0.0456)	(5.544)	(0.0560)	(5.445)

Notes: $^{***}P < 0.01$; $^{**}P < 0.05$; $^{*}P < 0.1$. Robust standard errors in parentheses. All models condition on a linear spline of the assignment variable, the teacher controls, and school-fixed effects.

Table A4 suggests less robustness to the performance gains associated with the HE threshold. For example, when the estimation sample is limited to the $n = 566$ observations within 20 points of the HE threshold, the estimated performance gains associated with HE incentives falls by more than half to 4.3 IMPACT points. However, because this sample reduction implies a loss of precision, it is also true that the 95 percent confidence interval on this point estimate includes the point estimate based on the full sample. The final shows that, in a nonparametric regression based on a triangular kernel, this RD estimate is somewhat larger and statistically significant.

Table A5 presents the results from four separate RD specifications that condition both on the actual thresholds that created contrasts in IMPACT incentives and on several false or "placebo" thresholds. In the case of the ME results, we see that the estimated retention and performance effects are larger and more precisely estimated at the threshold with real-world relevance and not at the placebo thresholds. These ad hoc specification checks (i.e., estimated effects concentrated where they should be and not where they should not) affirm the causal warrant of this RD design. However, it should be noted that, in the case of the HE results, this evidence is less dispositive. The far right column of Table A5 indicates that the estimated performance gains are at their largest for teachers at the actual HE threshold. However, in this saturated specification, this positive point estimate is noticeably smaller and falls short of statistical significance.