EdWorkingPaper No. 22-598

Incentives, Expectancy, and Social Identity: Field-Based Evidence from Teacher Evaluation

David Blazar
University of Maryland,
College Park

Melinda Adnot
University of North Carolina,
Charlotte

Xinyi Zhong
University of Washington,
Seattle

We examine heterogenous responses to job-embedded performance incentives along two dimensions theorized to drive motivation: (i) expectations of success when faced with easier versus more difficult tasks, and (ii) ones' social identity as part of a marginalized group (in our case, race). Compared to the largely lab-based literature on this topic, we leverage data from a fully implemented teacher evaluation system from the District of Columbia Public Schools over a seven-year period (2009-10 through 2015-16). Our regression discontinuity estimates reveal not only that task difficulty and social/racial identity drive much of the incentive effects, but also that there is a strong interaction between the two. Low-performing teachers threatened with dismissal improved much more on tasks with low difficulty and high expectations of success, relative to more difficult tasks (roughly 0.3 SD versus 0.15 SD). These trends were particularly pronounced for Black teachers, who experienced fewer successes than White teachers in the evaluation system generally. We also find that high-performing Black teachers were less responsive than White teachers to an incentive to increase their base pay. At the same time, the responses of Black teachers to the salary incentive were malleable and tracked closely with district-led redesign efforts that aimed to ensure greater equity in terms of teachers most likely to reap the benefits of this incentive.

VERSION: July 2022

INCENTIVES, EXPECTANCY, AND SOCIAL IDENTITY: FIELD-BASED EVIDENCE FROM TEACHER EVALUATION

David Blazar
University of Maryland, College Park

Melinda Adnot
University of North Carolina, Charlotte

Xinyi Zhong
University of Washington, Seattle

ABSTRACT

We examine heterogenous responses to job-embedded performance incentives along two dimensions theorized to drive motivation: (i) expectations of success when faced with easier versus more difficult tasks, and (ii) ones' social identity as part of a marginalized group (in our case, race). Compared to the largely lab-based literature on this topic, we leverage data from a fully implemented teacher evaluation system from the District of Columbia Public Schools over a seven-year period (2009-10 through 2015-16). Our regression discontinuity estimates reveal not only that task difficulty and social/racial identity drive much of the incentive effects, but also that there is a strong interaction between the two. Low-performing teachers threatened with dismissal improved much more on tasks with low difficulty and high expectations of success, relative to more difficult tasks (roughly 0.3 SD versus 0.15 SD). These trends were particularly pronounced for Black teachers, who experienced fewer successes than White teachers in the evaluation system generally. We also find that high-performing Black teachers were less responsive than White teachers to an incentive to increase their base pay. At the same time, the responses of Black teachers to the salary incentive were malleable and tracked closely with district-led redesign efforts that aimed to ensure greater equity in terms of teachers most likely to reap the benefits of this incentive.

JEL No. I21, J24, M52

This research is part of a research-practice partnership with the Strategic Education Research Partnership (SERP) Institute and the District of Columbia Public Schools (DCPS), with funding from the Institute of Education Sciences under grant R305H190057. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. We thank our DCPS partners in the Office of School Improvement and Supports and the Office of Teaching and Learning, who helped pose the research questions examined in this paper, provided access to the data, and gave feedback on analyses and drafts. We also thank our SERP partners for similar guidance, as well as for establishing the connection between university-based researchers and policy/practice-based partners. Correspondence regarding the paper can be sent to David Blazar at dblazar@umd.edu; Department of Teaching and Learning, Policy and Leadership, 2311 Benjamin Building, University of Maryland College of Education, College Park, MD, 20742.

1. Introduction

Contract and personnel economics are built on the fundamental premise that performance incentives motivate employees to change their behavior in order to maximize personal utility, ultimately increasing firm output (Lazear 2000; Holsmstrom 1979). A growing literature base provides support for this theory when performance incentives are implemented across job sectors (Weibel, Rost, and Osterloh 2010). At the same time, scholars point out how the theory can break down (Jacob and Levitt 2003) and have raised questions about the economic value. For example, merit pay for teachers is widely explored in both theoretical and empirical investigations on this topic (Borjas 2020; Ehrenberg and Smith 2016; Holmstrom and Milgrom 1991; Lazear 2003), and there is evidence of positive impacts on objective measures of teacher performance (Pham et al. 2021). At the same time, average effect sizes of merit pay schemes on the outcomes of teachers' students tend to be quite small (0.04 SD) relative to other performance-enhancing interventions for teachers that are similar in cost (upwards of 0.2 SD; Fryer 2017; Kraft, Blazar, and Hogan 2018).[1]

To better understand when, how, and for whom performance incentives work to modify employee behavior, economic models and analyses increasingly incorporate interdisciplinary theories

[1] Our back-of-the envelope calculations of the teacher incentive system under study in this paper come to a conservative estimate of roughly $4,000 per teacher per year (in 2022 dollars), which includes the following categories. First is the cost of performance monitoring primarily through classroom observations, at roughly $1,000 per teacher per year. On average, teachers were observed 3.4 to 4.9 times, depending on the school year. We assume 1.5 hours to prepare for, conduct, and debrief observations; and $100 per hour. We add the cost for hiring substitute teachers when teachers serve as the observers for other teachers. Taylor and Tyler (2012) estimate roughly $7,000 per teacher per year for performance monitoring and assistance. Second is the cost of the financial incentives, including both one-time bonuses and base salary increases. From the data, we know the exact amount of the bonuses, which comes to an average of roughly $2,300 per teacher per year. We also can observe whether or not teachers received a base salary increase, though the exact amount depends on teachers' starting salary; we estimate a per teacher and per year cost of $700. In a teacher evaluation system, merit pay for high-performers generally is paired with dismissal of low-performers, which include additional costs. However, here, we exclude costs associated with the hiring process because they are highly variable depending on the qualifications of outgoing/incoming teachers. We also note that some of the costs of hiring and of performance monitoring occurs in the absence of the incentives.

We compare these costs to those for professional development-focused interventions for teachers. For example, instructional coaching—in which an expert teacher works one-on-one with teachers to help identify areas for skill improvement and to develop these skills—is a highly effective intervention that results in increases in teacher practice (the main outcome type assessed in this paper) upwards of 0.5 SD and increases in student test scores of 0.2 SD (Kraft, Blazar, and Hogan 2018). Estimated costs for instructional coaching programs can vary across programs and contexts, but generally fall in the range of $4,000 to $10,000 per teacher per year (Kaufman et al. 2021; Knight and Skrtic 2021).

of motivation, drawing in particular from psychology and sociology (e.g., Gneezy, Meier, and Rey-Biel 2011). From psychology, expectancy theory posits that a key moderator of the effect of incentives on worker performance is one's expectation of success, often defined as the objective difficulty of the task and the proportion of individuals who complete that task successfully (Atkinson 1957; Locke and Latham 2002). Put simply, individuals are motivated to spend time and effort on the pursuit of activities on which they expect to do well. This noticing is intuitive but consequential to professions with multiple or complex tasks. For example, teachers are tasked with improving student achievement on standards-based assessments, as well as promoting curiosity and creative thinking, building interpersonal relationships with students, and managing the classroom environment. In this sort of setting, workers who are provided with an incentive must consider not just whether to change their behavior, but also where they should focus their attention amongst varied tasks and how they might improve in one or multiple tasks simultaneously (Holmstrom and Milgrom 1991).

Drawing from sociology, another likely factor shaping employees' behavioral responses to incentives is their social identity. Akerlof and Kranton (2000, 2005) initially theorized social identity broadly as one's sense of self relative to the organization, arguing that "outsiders" need larger monetary incentives than "insiders" to compensate them for acting in the interest of the firm rather than their own. Applications of this theory further consider social identity in reference to racial, ethnic, or income inequality, with some evidence suggesting that the power of performance incentives can be attenuated for groups that historically have been discriminated against and when that identity is made more salient (Farzana, Li, and Ren 2015; Hoff and Pandey 2006). One hypothesized explanation for this pattern is stereotype threat: if individuals from historically marginalized groups are expected to perform poorly, the stereotype may result in performance at that lower level (Steele, Spencer, and Aronson 2002). Another explanation is that experiences of marginalization among certain groups decrease expectations of success that, in turn, lead them to exert less effort. Relatedly, it could be the case that experiences of discrimination and marginalization—whether outside or inside the workplace—create a heightened need to focus on tasks with the greatest expectations of success. While some scholars have theorized links between expectancy and social identity (e.g., Lloyd and Mertens 2018), the empirical literature generally has treated them as separate.

Further, by and large, applications of expectancy theory and social identity to the traditional performance incentive model have been tested in researcher-defined scenarios rather than field settings (for reviews of the expectancy literature, see Locke and Latham 2002; Garber and Konradt 2014; Weibel, Rost, and Osterloh 2010). This pattern is especially pronounced for analyses of the moderating role of social identity where we are aware of just a couple of studies on this topic, all of which are simulations or lab-based (Farzana, Li, and Ren 2015; Hoff and Pandey 2006). Lab experiments make it easy to change the saliency of social identity in the incentive scheme, as well as to manipulate the difficulty of one or multiple tasks. However, lab environments may not parallel the field and on-the-job work because participants self-select into the experiment, are strictly scrutinized, and the stakes generally are low (Levitt and List 2007). In contrast, on-the-job performance incentives can create particularly high stakes—large changes in salary, as well as the possibility of dismissal—that are unlikely to be fully replicated in a lab or a simulation. The stakes may be even more pronounced for individuals or groups who have been marginalized outside or inside their work setting.

Using the case of performance incentives within the District of Columbia Public Schools' (DCPS) teacher evaluation system between the 2009-10 and 2015-16 school years, we make three contributions to the existing literature. First, our study is one of the few field-based analyses to confirm the largely lab-based literature on how individual responses to performance incentives can vary by task difficulty. To do so, we exploit strict eligibility thresholds in the teacher evaluation incentive structure, where the lowest-performing teachers were immediately dismissed, those just above them in the performance distribution were threatened with dismissal if they did not improve the following year, and the highest-performing teachers could earn large increases in their base salary (up to $27,000 per year) if they repeated their high performance the next school year. We pair this setup with rich performance measures derived from a classroom observation rubric that includes teaching tasks varying in their difficulty. On one end are tasks thought to be routine-oriented, guided by step-by-step procedures (e.g., maximizing instructional time). On the other end are non-routine tasks that generally are considered to be quite difficult to master (e.g., developing students' higher-level and conceptual understanding). Earlier analyses of the same DCPS data identify positive effects of dismissal threats and merit pay on an aggregate measure of teacher performance, on average across teachers (Dee, James, and Wyckoff 2021; Dee and Wyckoff 2015). However, analyses have not examined the moderating role of task difficulty or social identity, which we find shape much of the incentive effects.

Our regression discontinuity estimates indicate that dismissal threats resulted in improved performance on both the least and most difficult tasks, but that effects on the former were half as large as those on the latter (roughly 0.15 SD versus 0.3 SD). These findings align closely, for example, with those from a lab-based study by Shaw, Horton and Chen (2011), who created a short task in an online labor market where raters were asked to respond to closed-ended content analysis questions. Easy questions required simple yes/no responses, while the harder questions asked raters to identify a combination of answers from a pre-defined list. Raters who were presented at random with a financial penalty for incorrect responses outperformed those not presented with that disincentive, with the largest effects on the easiest tasks. The authors observed similar heterogeneity in effects by task difficulty when raters were offered a financial bonus, but these average effects were smaller compared to effects for the disincentive. We find similar patterns in a job that is substantially more complex and where the incentives match the high stakes often associated with full-time contracts.

Second, to our knowledge, we provide the first empirical evidence—whether in our outside of the lab—of an interaction between expectancy and social identity as moderating forces. The specific dimension of social identity that we explore is race because it has been especially salient in the DCPS context (DCPS 2022), and in teacher evaluation and incentive systems more broadly (Campbell 2020; Jiang and Sporte 2016; Steinberg and Sartain 2020). Although a majority of the DCPS teacher workforce is Black (see Table 1)—and so may be construed by some readers as "insiders" rather than "outsiders"—we show that Black teachers experienced the incentive and evaluation system quite differently from White teachers. Between the 2009-10 and 2015-16 school years, Black teachers were more likely than White teachers to be threatened with dismissal (3 percentage points) and less likely to be eligible for salary increases (5 percentage points). Black teachers also opted into the salary incentive—giving up some job protections—at substantially lower rates than White teachers (15 percentage points). DCPS leadership describe similar patterns in a recent equity memo (DCPS 2022). In a causal framework, we then find substantially smaller effects of the incentives on the subsequent performance of Black teachers versus White teachers. These differences were most pronounced for the salary incentives, where we observe null effects for Black teachers, on average, compared to substantively meaningful effects for White teachers (upwards of 0.3 SD). Further, while both Black and White teachers responded to the dismissal threats, the effects for Black teachers were concentrated in tasks with the greatest expectations of success, just as expectancy theory posits. For White teachers, effects of dismissal threats were similar in magnitude across teaching tasks.

In some ways, these patterns help extend prior lab-based analyses of the role of social identity in incentive schemes. For example, in a set of experiments in India, Hoff and Pandey (2006) assessed the number of puzzles solved by sixth and seventh graders when participants' caste was revealed versus when it was not; all participants received an incentive based on the number of puzzles completed. When caste was revealed, the performance of high-caste participants remained the same, whereas the number of puzzles completed amongst the low-caste participants declined by 20%. Farzana and colleagues (2015) found similar results in a lab-based experiment with primary-school students in China. Similar to these studies, our analyses reveal that salary incentives are less impactful for Black teachers relative to their White colleagues, but in a real-world context where the consequences of the incentives are much higher.

We further argue that our findings support an expectancy-based hypothesis for the differential responses of Black versus White teachers, rather than one related to stereotype threat. A key distinction between the prior literature and ours is that the lab experiments randomly assigned whether social identity was or was not revealed. This setup leads Hoff and Pandey (2006) and Farzana and colleagues (2015) to interpret their findings through a stereotype threat framework, arguing that publicly revealing the identity of individuals from historically marginalized groups can increase negative thoughts about oneself and decrease confidence, which in turn impedes success in an incentive system. In contrast, in our context, teachers' race always was revealed and salient, meaning that externally imposed stereotypes always were present. Given that the responses to dismissal threats on tasks with the greatest expectations of success were very similar for both Black and White teachers, it seems unlikely that stereotype threat was the prevailing force for differential responses between these two groups on other tasks with the least expectation of success.

Further, in a set of exploratory analyses, we find that the responses of Black teachers to the salary incentives tracked closely with district-led redesign efforts that aimed, at least in part, to address concerns about which teachers were most likely to reap the benefits of this incentive. In other words, the effects of the salary incentives appear to be malleable and potentially linked to forces that influence expectancy. In the first year of the evaluation system, Black teachers responded to the salary incentives, with particularly large effects (upwards of 0.6 SD) once accounting for take-up of the incentive offer. However, salary incentive effects for Black teachers fell to zero in the second and third years of implementation, aligning closely to the historical record documenting community protests and concerns that the evaluation system was not serving Black teachers' best interests (Cardoza 2011;

Whitmire 2011). Then, following a substantial district-led redesign that restricted eligibility for the salary incentive to teachers working in high-poverty schools where Black teachers were overrepresented, the Black-White gap in opt-in rates for this incentive shrank substantially (to a low of 3 percentage points). And, estimates for the effect of salary incentives for Black teachers turned positive (though do not rise to the level of statistical significance). We do not claim any causal relationship between redesign efforts and incentive effects. Rather, we argue that fluctuations in the responsiveness to salary incentives among Black teachers likely are indicative of shifting expectations of success, rather than other forces such as stereotype threat.

More narrowly, our findings contribute to the teacher incentives literature. Merit pay for teachers has been discussed and implemented in school systems for many decades (e.g., Moore-Johnson 1984; Murnane and Cohen 1985), but has gained substantial attention particularly in the U.S. and in the past 15 years. For example, under the Obama administration's Race to the Top initiative rolled out in 2010, states were able to receive large grants for implementing teacher evaluation systems that attached performance metrics to job decisions (McGuinn 2012). Within this period, all but six U.S. states implemented some form of teacher evaluation (Bleiberg et al. 2021) that incentivized improved performance through both rewards and sanctions (Howell and Magazinnik 2017). Because elementary and secondary schools are one of the largest employers in the U.S. (Bureau of Labor Statistics 2020), widespread use of incentives for teachers represents a critical and consequential case of contract design, with implications for other sectors as well.

Our findings indicate that job incentives can work to improve teacher performance, but that the effects are not universal. The fact that we find *any* effect of the incentives on teacher outcomes is encouraging, given earlier concern that the very nature of teaching made it ill-suited for improvement through incentives (Murnane and Cohen 1986). Teaching is a multitask profession without consensus on the specific combination of tasks that are needed to achieve desired outcomes, namely student learning. Similar to our findings—which do identify positive effects of incentives—a recent meta-analysis of the impact of the merit pay component of teacher evaluation also found benefits on student test-score performance (Pham et al. 2021). At the same time, the magnitude of the effects on student outcomes documented in the meta analysis (0.04 student-level SD) is small relative to other approaches for improving teacher performance such as instructional coaching in which experts observe and provide feedback on teachers' practice (upwards of 0.2 SD; Fryer 2017; Kraft, Blazar, and Hogan 2018). From our analyses, heterogenous effects by task difficulty and teacher race further suggest that the theory undergirding teacher evaluation incentives is not working fully as intended.

In our opinion, heterogenous effects between Black and White teachers are particularly important for school systems to consider when designing and implementing performance-based job incentives. Black teachers are underrepresented in the U.S. compared to populations of Black students (Schaeffer 2021). These mismatches are particularly troubling given the very large effects of Black teachers on the test scores of Black students (0.2 SD; Dee 2004) and on the intrapersonal skills of all students (0.3 SD; Blazar 2021). Our results indicate differences in the responses of Black versus White teachers to the salary increases and, to a lesser extent, dismissal threats. As such, we argue that contract, evaluation system, and incentive system design must consider the responses of Black teachers, particularly when a growing policy goal across U.S. states is to recruit and retain more of them (DeRamus-Byers 2021; Education Commission of the States 2019).

2. Performance Incentives for Teachers in the District of Columbia Public Schools

DCPS introduced the IMPACT teacher evaluation system in the 2009-10 school year with a central goal of shifting the average quality of the teacher workforce through two incentive-based mechanisms: dismissal threats and salary incentives. Both incentives are aligned to assumptions of rational, utility-maximizing employees who aim to keep their job and earn the highest possible salary

(Lazear 2000; Holmstrom 1979). Like other teacher evaluation systems (Bleiberg et al. 2021), IMPACT includes mechanisms other than "carrot and stick" incentives as means of shifting system-wide teacher quality and student outcomes. The evaluation system provided an avenue for clarifying a vision of excellent instruction and providing feedback to teachers about how to improve their practice (Phipps and Wiseman 2021). The district's theory of change also posits benefits for students stemming from the redistribution of teachers, including replacing low-performing teachers with higher-performing ones and incentivizing high-performing teachers to the move to the highest-needs schools (Adnot, Dee, Katz, and Wyckoff 2017). In this paper, though, we focus only on the effects of dismissal threats and salary increases.

We also focus on the earliest years of the incentive system—through the 2015-16 school year—even though the system continues to operate today. In the 2016-17 school year, the district adopted a new rubric to monitor teacher performance, creating uncertainty about whether possible shifts in teacher responses reflect incentive effects or learning of a new rubric. Further, our analyses aim to document behavioral responses to incentives using DCPS as a case, rather than evaluating the IMPACT system specifically, where the most recent data may be more relevant.

2.1. Monitoring and Measuring Teacher Performance

To determine eligibility for and receipt of incentives, teacher performance was monitored on a yearly basis through a multiple-measures system, including: (i) observations of classroom instruction scored on a standards-based rubric (up to 75% of the total score, depending on the availability of other metrics); (ii) student achievement growth on a district-administered assessment (up to 50% for teachers who worked in a grade and subject mandated for high-stakes testing); (iii) student achievement on a teacher-selected assessment (up to 15%); (iv) principals' assessment of teachers' commitment to the school community (up to 10%); and (v) school-aggregated student test-score performance (up to 5%, but only in the first three years of the evaluation system). The summary measure of teacher performance—which was a weighted average of the individual components—ranged from 100 to 400, with multiple bands and thresholds that determined the allocation of incentives (see Figure 1). At inception, there were four performance bands: "Ineffective", "Minimally Effective", "Effective", and "Highly Effective". In the 2012-13 school year, the middle band was split in two ("Developing" versus "Effective").

Performance monitoring vis-à-vis teachers' contributions to student test-score growth—often referred to as teacher "value-added"—was a district-level priority in the design of the evaluation system (Dee and Wyckoff 2015; Whitmire 2011). In practice, though, only 15% of teachers worked in a grade and subject where district-wide testing was required (i.e., math and English language arts in grade 3 through 8 and once in high school; see Table 1). When teacher value-added was not available, other metrics received greater weight. On average across teachers, scores generated from observations of classrooms accounted for 69% of the overall score; for over 90% of teachers, observations accounted for over 50% of their performance score.

In our analyses, we focus on classroom observations as the primary measure of teacher performance for two reasons. First, because these observation scores comprised the majority of the total evaluation score for the majority of teachers, it is intuitive that teachers' behavioral responses would focus on tasks and skills identified in the rubric. Second, the teaching tasks and skills being monitored capture a wide range of classroom practices that vary substantially in terms of difficulty, thus facilitating analyses related to expectancy. The district-created rubric used to score the quality of teachers' classroom instruction was known as the Teaching and Learning Framework (TLF) and drew from instructional research from several other observation instruments (Danielson 2011; Pianta and Hamre 2009; Wiggins and McTighe 2005), with goals of creating a common language to discuss teaching and learning and for providing clear expectations for teacher performance (DCPS 2010). The framework was originally designed by teachers, school leaders, and central office staff in 2008-09 and was streamlined after the first year of implementation, yielding nine total components (see Table 2).

Aligned to Atkinson's (1957) definition of objective task difficulty, we organize and discuss the nine components based on average performance across all DCPS teachers. The most difficult tasks asked teachers to develop students' higher-level understanding through effective questioning and by making sense of concepts rather than solving procedures, as well as to engage students across all learning levels in rigorous content. Average scores on these two dimensions were below 3 out of the 4-point scale (2.72 for Higher-Level Understanding, and 2.95 for Rigorous Work). The least difficult performance metrics included the extent to which teachers Maximized Instructional Time (average score of 3.3 out of 4) and Built a Supportive, Learning-Focused Classroom (3.4 out of 4). These patterns are consistent with other research examining differences in teaching practices, with metrics relating to questioning technique often receiving the lowest scores and metrics related to the classroom environment often receiving the highest scores (Garrett and Steinberg 2015; Hamre et al. 2013; Kane, Taylor, Tyler, and Wooten 2011; Kane and Staiger 2012). The rank order of tasks based on average teacher scores are almost identical when disaggregating by teacher race or school year.[2]

Depending on their prior-year evaluation score, teachers were observed up to five times each year: three times by an administrator from teacher's own school (often the principal) and twice by a content area expert employed by the district expressly for the purpose of conducting evaluations (called a master educator). Seventy-five percent of teachers in our data were observed at least four times per year. In the evaluation process, teacher scores were averaged across all observations within

[2] Ideally, we would determine task difficulty based on the first year of performance data, which is never used as an outcome measure in our analyses. However, as noted elsewhere, the observation rubric was revised between the first and second years of implementation. Of those tasks that showed up in both versions of the rubric, rank ordering is the same. We are confident in our identification the most versus least difficult tasks given consistency of patterns in the DCPS data across years and groups of teachers, as well as alignment to patterns in other datasets that include classroom observation scores using different rubrics but with overlapping teaching tasks (Garrett and Steinberg 2015; Hamre et al. 2013; Kane, Taylor, Tyler, and Wooten 2011; Kane and Staiger 2012 Garrett and Steinberg 2015; Hamre et al. 2013; Kane, Taylor, Tyler, and Wooten 2011; Kane and Staiger 2012).

a given school year, which aimed to minimize measurement error due to the foci of a given lesson, differences in scoring across raters, etc. (Hill, Charalambous, and Kraft 2012; Kane and Staiger 2012). Similarly, in our primary analyses, we use observation scores averaged across all available lessons. To decrease measurement error, we follow a generalizability framework (Hill, Charalambous, and Kraft 2012) and shrink observation scores back to the mean based on the total number of lessons per teacher. In a robustness test, we re-estimate results using performance scores captured by master educators, who may have had less of a stake than school leaders in how teachers performed and whether or not they received a given incentive. Unlike school leaders, master educators also were not privy to information about a given teacher's prior performance (Dee, James, and Wyckoff 2021).

Psychometric analyses of these data indicate that the multiple-lesson, multiple-rater observation process was successful in capturing differences in teacher rather than rater behavior. In Table 2, we show that the amount of variation at the teacher-year level—as opposed to construct-irrelevant sources of variation such as raters—is at or above 63% for each of the nine dimensions, which is similar to statistics generated from other datasets (Hill, Blazar, and Lynch 2015; Kane and Staiger 2012). Comparatively, we observe roughly 10% of the variation at the rater level. Important for both policy and research, these rater-level differences do not appear to be driven by systematic racial biases. The amount of variation at the rater-by-teacher race level is no higher than 5%.

2.2. The Incentive Structure

Teachers who scored at the lowest end of the performance distribution (i.e., "Ineffective"; see Figure 1) were immediately dismissed, and those who scored in the second-lowest category (i.e., "Minimally Effective") were threatened with dismissal if they did not move up the performance distribution in the next school year. In our analyses, we compare teachers who received a dismissal threat to those teachers who just barely missed it because they scored in the next-highest performance band (i.e., "Effective" in the first three years of implementation, and "Developing" in subsequent years when the "Effective" category was split in two). As shown in Table 3, for the years of the system under study in this paper, roughly 3% of all teachers received an "Ineffective" rating meant to lead to immediate dismissal, and 8% of all teachers received a "Minimally Effective" rating that resulted in a dismissal threat. In Table 3, we highlight rows in different shades of gray to align with different periods of evaluation system (re)design, which we describe below. Of those teachers who were threatened with dismissal, 25% were dismissed the next school year. That DCPS actually dismissed low-performing teachers stands in sharp contrast to most other school districts' teacher evaluation systems, where almost everyone receives a satisfactory rating (Kraft and Gilmour 2017; Weisberg et al. 2009).

Rates of dismissal and dismissal threats often were higher for Black teachers relative to White teachers, with gaps as large as 5 percentage points (see Table 3). These trends were related, in part, to the overrepresentation of Black teachers in high-poverty schools, defined by DCPS as schools where 60% or more of students were eligible for free or reduced-price meals (see Appendix Table 1). On average across school years, 11% of teachers in high-poverty schools were threatened with dismissal, compared to 3% of teachers in low-poverty schools. In high-poverty schools, 56% of teachers were Black and 25% were White, compared to low-poverty schools where 30% of teachers were Black and 54% were White. Other research shows that high poverty rates within schools can create environments less conducive to high-quality teaching (Boyd et al. 2008; Sass et al. 2012).

At the top end of the performance distribution, teachers who earned the highest rating (i.e., "Highly Effective") received an offer to opt into IMPACT*plus*, which made them eligible to receive an immediate, one-time bonus of up to $25,000[3], as well as a permanent increase in base pay if they received this same rating the following school year. Base-pay increases, which are the focus of our

[3] The exact amount of the one-time bonus depended on: (i) the poverty rate of the school; (ii) the test-score performance of the school; (iii) whether or not the teacher worked in a high-needs subject area; and (iv) whether or not the teacher worked in a grade level and subject area where students took district-administered tests, meaning that one metric of performance was the teacher's contribution to test-score growth. For details, see Dee, James, and Wyckoff (2021).

analyses, started at roughly $7,000 and could be as large as $27,000 depending on teachers' years of experience in the job and their education level.[4] Education and experience determine base salary in DCPS and in most other school districts across the U.S. (Hanushek 2007). Opting into this portion of the evaluation system required that teachers give up their contractual right to look for a new job for a year without losing pay or benefits, if they lost their current teaching position. Teachers received the opt-in offer in the spring of the school year in which they earned the high-performance rating, and they could not reverse this decision in later years. Across school years examined in this analysis, roughly two-thirds of eligible teachers scoring "Highly Effective" opted into IMPACT*plus* (see Table 3). We focus on the base-pay incentive in our analyses and not the one-time bonus, because the need for repeated high-performance provides a useful discontinuity between those teachers who just missed and those just met the threshold that determined eligibility for the salary incentive.

Following similar patterns for race-based gaps in dismissal threats, 8% of Black teachers were offered a salary increase, compared to 13% of White teachers (see Table 3). These trends may also be related to the overrepresentation of Black teachers in high-poverty schools, where 17% of teachers received a "Highly Effective" rating compared to 44% of teachers from low-poverty schools (see Appendix Table 1). We also observe a large gap in the opt-in rates to IMPACT*plus* of Black versus White teachers, with Black teachers less likely to opt in by 15 percentage points (see Table 3).

2.3. Incentive System Redesign and Responses to Controversy

Over the seven-year period under investigation in this study (2009-10 through 2015-16 school year), DCPS made several changes to the IMPACT evaluation system design and the incentive

[4] In the first three years of implementation (i.e. 2009-10 to 2011-12 school years), teachers who repeated their high performance across two consecutive years and worked in a school where 60% or more of students were eligible for free or reduced-price meals received five "service credits", generally equivalent to years on the job; they also moved from the bachelor's to master's degree band, if they were not already is this band. Teachers who worked in schools where less than 60% of students were eligible for free or reduced-price meals moved to same amount in terms of degree-equivalence, but received three rather five years of service credits. Starting in the 2012-13 school year, only those teachers who worked in high-poverty schools were eligible for an increase in base-pay, with the same five-year service credit offering.

structure. As noted above, the classroom observation rubric was revised slightly for the second year of implementation (i.e., 2010-11). Starting in the fourth year of implementation (i.e., 2012-13), the middle performance band was split in two (see Figure 1), which aligned with the rollout of additional incentives. Beginning this year, teachers were threatened with dismissal for multiple combinations of low scores: one "Ineffective" rating (same as in prior years), two consecutive "Minimally Effective" ratings (also the same as in prior years, and still the threshold that we exploit in our analyses), one "Developing" rating followed by one "Minimally Effective" rating, or three consecutive ratings below "Effective".[5] The revised system also created a multi-tier career ladder with additional opportunities for financial rewards. The largest financial rewards still required teachers to repeat their "Highly Effective" performance over two consecutive school years (a fact that we leverage in our empirical strategy), while "Effective" teachers could receive smaller increases in base pay as they moved up the performance distribution on a year-to-year basis. Also starting in the 2012-13 school year, the weights for teachers' summative evaluation score were adjusted, decreasing the emphasis on teachers' contributions to student test-score growth (where applicable) and increasing the emphasis on classroom observations. Finally, all financial rewards were limited to teachers in high-poverty schools, meaning that 60% or more of students were eligible for free or reduced-price meals.

These adjustments inform sample restrictions that we describe below, but more importantly illustrate contextual factors that we hypothesize may inform expectations of success within the system, particularly for Black teachers. To implement the evaluation system and its incentives, buy-in was needed from multiple stakeholders including the Washington D.C. mayor, DCPS chancellor of

[5] We focus only on the dismissal threat that teachers faced for having two consecutive "Minimally Effective" ratings, as this is a rule that applies across all school years. Adding new routes to dismissal does not alter the sharp discontinuity between receiving one "Minimally Effective" rating versus scoring in the next highest performance band. However, the introduction of more avenues to dismissal changed the incentive contrast to some extent from the teacher perspective. As Dee, James, and Wyckoff (2021) write: "the Minimally Effective"/"Developing" contrast effectively compares the credible and immediate dismissal threat for "Minimally Effective" teachers who d[id] not improve immediately to the incentives faced by "Developing"-rated teachers who instead ha[d] two years to achieve an "Effective" rating" (p. 315). When we disaggregate dismissal threat effects by year (i.e., prior to versus after this change), we find fairly similar results.

schools, the teachers' union, city councilmembers, and others. However, broad buy-in was short-lived. Following the initial round of teacher firings in spring 2010, teachers and community members—particularly Black individuals—raised concerns at community meetings and through protests. In fall 2010, the mayor who appointed the DCPS chancellor who then established the teacher evaluation system lost his bid for reelection in the Democratic primary, driven largely by shifts in voting blocs in majority Black wards in the city; exit polling indicated that education was a primary reason for this shift (Whitmire 2011). During this same time, the local teachers' union sued the district as a means of challenging teachers' low ratings and dismissals.[6] While the lawsuit did not focus on teacher race, public commentary did. While discussing the lawsuit, the president of the local union referred to the evaluation system as "unjust" due to a "context of racis[m]" and "discriminat[ion]" (Cardoza 2011).

Despite union and some community pressure to do so, DCPS leadership did not get rid of the teacher evaluation system, which still exists today. However, redesign efforts aimed, at least in part, to address concerns of inequitable allocation of incentives along racial and other lines. In particular, limiting the salary incentives to teachers who worked in high-poverty schools in the 2012-13 school year shifted attention towards Black teachers who were overrepresented in these schools (see Appendix Table 1). Prior to this shift, the proportion of Black teachers offered the salary incentive was roughly half the proportion of White teachers also offered the incentive (see Table 3). Twelve percent of Black teachers were offered the salary incentive in the 2009-10 and 2011-12 school years, and 4% were offered the incentive in the 2010-11 school year. (The drop in the 2010-11 school year is mechanical, given that teachers could receive an incentive only once in a two-year period due to the requirement for two consecutive years of high performance.) In contrast, 22% of White teachers were offered the salary incentive in the 2009-10 school year, 8% in the 2010-11 school year, and 24% in the

[6] Washington Teachers' Union, Local #6, American Federation of Teachers, AFL-CIO v. District of Columbia Public Schools, App. 11-CV-1104 (2010).

2011-12 school year. Following the 2012-13 redesign, the proportion of Black and White teachers who received a salary incentive offer were almost identical (i.e., 5% to 10% for each group).

Changes in the evaluation system design and incentive structure also correlate with fluctuations in opt-in rates to the salary incentive that required teachers to give up some of their job protections. For both Black and White teachers, opt-in rates were lowest during periods of uncertainty (i.e., in the first year of implementation of the evaluation system in 2009-10, and in the first year of the 2012-13 redesign). In these early years, the gaps in opt-in rates between Black and White teachers also were quite large, upwards of 20 percentage points in the same two periods of heightened uncertainty. Even when overall opt-in rates increased across the board in the second and third years of implementation, Black teachers were roughly 15 percentage points less likely than White teachers to opt into the salary incentive. However, following the decision to restrict salary incentives to teachers in high-poverty schools, the gap in opt-in rates between Black and White teachers shrank substantially to a low of 3 percentage points in the 2014-15 school year. We interpret these shifts as evidence of shifting expectations of success, particularly for Black teachers.

3. Empirical Strategy, Data, and Sample

To estimate the effect of dismissal threats and salary increases on subsequent teacher performance, we exploit the sharp incentive contrast that teachers experienced based on their overall evaluation score. A teacher who scored 249 on the summative 100- to 400-point IMPACT scale is assumed to be no different than a teacher scoring 250, except one teacher received a low-performance signal (i.e., “Minimally Effective”) and threat of dismissal in the next year if she did not improve, while the other received the message that her performance met the district’s standard. A similar discontinuity exists at the high end of the performance distribution, where teachers who scored 350 (i.e., “Highly Effective”) were eligible for a large salary increase the following year, while teachers who scored 349 were not. As documented elsewhere (Dee and Wyckoff 2015; Dee, James, and Wyckoff 2021) and available upon request, in the years of data used in this analysis, evaluation scores perfectly predicted performance bands and the incentives associated with them.[7]

The core estimating equation for our regression discontinuity (RD) design is as follows:

$$Y_{is(t+1)}^{m} = \alpha + \beta I(S_{it} \leq 0) + f(S_{it}) + \pi_t + \gamma X_{it} + \delta_s + \varepsilon_{it} \qquad (1)$$

where Y is a measure of teaching practice on task *m* for teacher *i* in school *s* and year *t*+1. We capture outcomes a year after teachers received (or did not receive) the incentive, as both the dismissal threats and salary incentives required repeated performance across two consecutive years. The parameter β reports the effect of receiving an initial "Minimally Effective" or "Highly Effective" rating on a given teaching task, or the "jump" in the outcome variable at the relevant performance threshold, conditional on a centered function of the initial evaluation score, S_{it}, that serves as the forcing variable. In our primary analyses, we pool data across all school years, and so include year fixed effects, π_t. In some models, we further condition on teacher covariates, X_{it}, and school fixed effects, δ_s; ε_{it} is a mean-zero error term, and robust standard errors are reported to allow for heteroskedasticity.

Whereas other scholars report positive effects of the DCPS job incentives on an aggregate measure of performance and on average across teachers (Dee and Wyckoff 2015; Dee, James, and Wyckoff 2021), we are interested in differential responses to the incentives along two dimensions: task difficulty and teacher race. To examine heterogeneity by task difficulty, we start by estimating equation (1) for a subset of dimensions of classroom practice identified in Table 2 as the most versus least challenging. For the sake of parsimony, we focus on the two most difficult and the two least difficult dimensions, which reflect tasks with substantially different expectations of success. We formally test

[7] In DCPS, teachers were allowed to appeal their evaluation score, which could introduce bias into our estimates. To avoid this possibility, we use teachers' initial score to determine assignment to treatment. In practice, appealing scores was quite rare. In the first year of implementation, 1.75% of all teachers appealed their score and 0.05% of teachers had their score changed. After that, appeals and changes occurred for no more than 0.5% of teachers.

for differences in the magnitude of the incentive effects across tasks by combining the variance-covariance matrices for each analysis and performing Wald tests of coefficient equivalence. For analyses examining differential responses by teacher race, we interact the treatment indicators and the function of the forcing variable in equation (1) with dummy variables identifying Black teachers and White teachers. We focus on these two groups only, as teachers from other racial/ethnic backgrounds comprised no more than 4% of the DCPS teacher workforce (Table 1). In the interacted models, we estimate effects of the incentives for Black and for White teachers, and then test the equality of coefficients. Patterns of results are the same if we run models separately for subgroups of teachers.

To identify our analytic samples, we start with the full population of DCPS teachers—roughly 3,500 individuals per year—from the 2009-10 to 2014-15 school year.[8] We include data from the 2015-16 school year, but only for capturing outcomes triggered by incentive eligibility in the prior year. Next, we define the dismissal threat and salary incentive samples by specifying a maximum bandwidth of 50 points on either side of the eligibility thresholds. Although we could include wider bands in some school years (see Figure 1), this limitation ensures consistency across years. A maximum bandwidth of 50 also is appropriate in the 2009-10 through 2011-12 period, when there was only one performance band separating the "Minimally Effective" and "Highly Effective" teachers, with each of these two groups eligible for different incentives. Our restriction to a 50-point bandwidth means that teachers in the middle band (i.e., "Effective") serve as the comparison group for just one of the incentivized groups.

For the salary incentive analysis sample, we further exclude teachers who received a salary increase the prior year. Teachers were able to receive multiple salary increases, but the requirement for two consecutive "Highly Effective" ratings reset each time. In 2012-13 onward, we also exclude

[8] We exclude a small fraction of teachers who were rated on an observation instrument other than TLF, which generates our primary out measures. Similarly, we exclude full schools—and all teachers within them—devoted to supporting students with special education needs, as these schools often used a classroom observation rubric other than TLF.

teachers from the salary incentive sample if they worked in schools where fewer than 60% of students were eligible for free or reduced-price meals, as these teachers were not eligible for salary increases. Starting in 2012-13, the introduction of a career ladder made teachers rated both as “Highly Effective” and “Effective” eligible for salary increases (the one for the latter group being much smaller than the one for the former). However, once teachers reached the middle rung of the career ladder, the only salary incentive available to them was the one that required two consecutive “Highly Effective” ratings. Therefore, for these later years, we limit the salary incentive analysis sample to those “Effective” and “Highly Effective” teachers above the middle rung of the career ladder, where only the “Highly Effective” teachers were eligible for a base salary increase. These sample restrictions echo the technique of frontier RD, which uses multiple variables to determine assignment to treatment (Reardon and Robinson 2012; Wong, Steiner, and Cook 2013).

Finally, following Dee and colleagues (2021), for the dismissal threat sample, we exclude the 2009-10 school year given anecdotal discussion with DCPS leadership and empirical evidence that the dismissal incentive was not yet fully implemented (Dee and Wyckoff 2015). In that year, a summative score below 250 was not a perfect predictor of a “Minimally Effective” rating. Comparatively, summative scores perfectly predicted ratings that triggered dismissal threats in all subsequent years. Summative scores also perfectly predicted ratings that triggered salary incentives in all years.

RD designs have a strong causal warrant due to the arguably random assignment of treatment right around sharp cut points (Campbell 1969; Lee and Lemieux 2009). An important concern with any RD design is that there may be systematic sorting across the performance threshold. If teachers were able to manipulate the variable that “assigned” them to one side of the threshold or the other, this introduces bias into the estimates because there are likely other differences between teachers. Literature on RD designs recommends a number of analyses to provide a check on this assumption (Imbens and Lemieux 2008; Lee and Lemieux 2009; McCrary 2008), and empirical examination in our data suggests that the assumption holds. In Table 4, we show estimates from regression models that predict treatment assignment as a function of distance from the threshold, year fixed effects, and observable background characteristics of teachers and their teaching positions. For both the dismissal threat and salary incentives samples, we cannot reject the null hypothesis of joint differences in the characteristics of teachers just passing versus just missing incentive eligibility (*p* = 0.186 for the dismissal threat sample, and 0.182 for the salary incentive sample). McCrary tests, which examine the density of observations around the performance thresholds, also fail to reject the null hypothesis of a smooth distribution across the eligibility threshold (*p* = 0.934 for the dismissal threat sample, and 0.456 for the salary incentive sample).

It is important to note that there is only outcome data in our sample for teachers who returned to an instructional position in the next year (see Table 1). This form of sample attrition is not an internal-validity threat but rather a component of the evaluation system theory of change, which incorporates both incentive and de-selection effects. Dee and Wyckoff (2015) aim to tease out one from the other by estimating an RD specification where summative evaluation performance in the prior year is the dependent variable, finding "small and statistically insignificant effects that are consistent with the hypothesis of behavioral change in response to the incentives" (2015. p. 21). Relatedly, in Table 4, we examine whether teachers in the treated and untreated groups are balanced on observable characteristics across the eligibility threshold for those who remained in the district the following year, finding no differences (*p* = 0.176 for the dismissal threat sample, and 0.601 for the salary incentive sample). In others words, even though some teachers voluntarily left the district after learning their rating, the background characteristics of teachers of who remained were not significantly altered. One important implication of this is that our estimated effects should be interpreted as improvements in practice conditional on teachers' decisions to remain in the school system.

4. Results

To begin, in Figure 2, we show graphical evidence of the impact of dismissal threats (Panel A) and financial incentives (Panel B) on teachers' subsequent observation scores on the classroom observation rubric (TLF). Here, the score is an average of all nine teaching tasks. We also pool data across teachers from all racial and ethnic backgrounds. Visually, the graphs show jumps in the aggregate measure of classroom practice both for teachers who received a dismissal threat (0.24 SD) and those who received a salary incentive (0.08 SD). We focus here on the aggregate teacher observation score because dimensions of teaching practices are the focus of subsequent analyses. Average incentive effects are similar when focusing instead on the summative IMPACT rating that includes observations, student-test score growth, and other metrics (see Table 5). One point of comparison for interpreting these effect sizes is the difference in average classroom observation scores between veteran teachers and those in their first three years of experience (0.35 SD). On their own, these findings suggest that the incentives worked to improve subsequent teacher performance.

At the same time, as we describe in detail in the following section, these average effects mask substantial heterogeneity by specific instructional task, teacher race, and time period of evaluation system (re)design.

4.1. Heterogeneous Incentive Impacts

In Table 5, we report estimates of the effect of dismissal threats and salary incentives on six performance metrics: the overall IMPACT evaluation score, the aggregate measure of classroom performance referenced in Figure 2, the two most difficult tasks with the lowest expectations of success, and the two least difficult tasks with the greatest expectations of success. We both pool data for teachers across racial/ethnic backgrounds and then disaggregate results for Black versus White teachers. For the pooled sample of all teachers, each estimate comes from a separate regression model that controls for year fixed effects and a linear function of the forcing variable. As described below, estimates are qualitatively similar when we further control for background teacher and school characteristics, when we replace school characteristics with school fixed effects, and when we specify a quadratic function of the forcing variable. Estimates by teacher race come from models that include Black and White teachers only. All outcome measures are standardized within school year to have a mean of 0 and a SD of 1, and so estimates can be interpreted as standardized effect sizes.

For all teachers and most tasks, dismissal threats led to substantial performance improvements in the following school year. Consistent with an expectancy framework, in the pooled sample of all teachers, effects on the least difficult tasks with the greatest expectations of success (roughly 0.30 SD) are twice as large as effects on the most difficult tasks with the lowest expectations of success (roughly 0.15 SD). *P*-values of coefficient equivalence shown at the bottom of the table indicate that the magnitudes of these effects are statistically significantly different from each other. We observe very similar patterns when disaggregating effects for Black teachers, which makes sense given that Black teachers comprised a disproportionate share of teachers in the dismissal threat analysis sample (61%) relative to the larger district population (50%; see Table 1). In comparison, when disaggregating effects for White teachers, we cannot reject the null hypothesis of no difference in effect sizes across teaching tasks that vary in terms of their difficulty. The effects of dismissal threats on the subsequent performance of White teachers are large (upwards of 0.52 SD). Comparing effects of dismissal threats for White versus Black teachers, we can reject the null hypothesis of a difference in the magnitude of effects for one of the four classroom observation tasks with the lowest expectations of success (i.e., Higher-Level Understanding), as well as for the two overall performance measures.

Turning next to the effects of salary incentives, between-group differences are more pronounced. For Black teachers, salary incentives had no statistically significant impact on subsequent performance on any metric. Point estimates are right around zero, and sometimes negative. Comparatively, for White teachers, salary incentives led to improvements in subsequent performance upwards of 0.3 SD. On the measures where we see salary incentives effects for White teachers, we can further detect statistically significant differences relative to impacts for Black teachers. Examining salary incentive effects by task difficulty, patterns for White teachers are inconsistent with an expectancy framework: effects are largest for the most difficult tasks relative to the least difficult ones, and these differences are statistically significant. One possible explanation may be that the difficult tasks provided the most room for improvement; comparatively, high-performing teachers eligible for salary incentives already were excelling at the least difficult tasks (see Table 2). However, as we describe below, for Black teachers eligible for salary incentives who also scored at the top end of the scale for the least difficult tasks, we do see some improvements in these areas in some school years. Our estimates of the effect of salary incentives—as well as those for dismissal threats—are reasonably precise with standard errors around 0.04 SD, showing that our research design has the power to detect small effects and those that are of substantive significance.

To further probe the stark differential responses of Black versus White teachers to the salary incentives, we further disaggregate effects by time periods in order to examine whether group-specific trends correspond with system-level attention to expanding the set of teachers reaping the benefits. In Table 6, we focus on three periods: (i) the first year of implementation (2009-10 school year); (ii) the next two years of implementation, after teachers had experienced the system for a year and when there were widespread community protests (2010-11 and 2011-12); and (iii) the following three years of implementation after the salary incentives were restricted to teachers in high-poverty schools where Black teachers were overrepresented (2012-13 through 2014-15). When disaggregating effects by time period, statistical power expectedly decreases, though we still detect several notable patterns.

Across the three time periods, we observe consistent, positive effects of the salary incentive for White teachers. There are some differences across years in terms of the specific tasks where White teachers improved most, though the differences generally are not large. For Black teachers, disaggregation of salary incentive effects by time periods reveals large fluctuations. In the first year, we estimate salary incentive effects for Black teachers upwards of 0.23 SD.[9] In comparison, in the second and third years of implementation, salary incentive effects for Black teachers are negative for all but one metric (but generally small and not statistically significantly different from zero). With the shift in evaluation system design in the fourth year of implementation, point estimates for Black teachers all turn positive (but do not rise to the level of statistical significance).

In Appendix Table 2, we conduct the same disaggregation by year for the effects of dismissal threats, finding fairly consistent estimates. This is unsurprising, given that the redesign efforts focused primarily on changing the salary incentive structure. In this appendix table, we show estimates for two time periods, excluding the first year when the dismissal threats had not been fully implemented (see Dee, James, and Wyckoff 2021; Dee and Wyckoff 2015). Small differences in dismissal threat effects between the two time periods potentially reflect the fact that, starting in the 2012-13 school year, additional sets of teachers were eligible for dismissal with different combinations of low scores (see discussion above on changes to the incentive structure). Here, we compare teachers who received an immediate dismissal threat to those who also were subject to dismissal but had more time to improve. This contrast is not as strong as in earlier years, where we compare teachers who received an immediate dismissal threat to those teachers who did not receive any threat.

4.2. Take-Up of Salary Incentive Offer

Do the between-group differences in the effect of salary incentives simply reflect differences in opt-in rates? As a reminder, teachers who scored at the top end of the performance distribution

[9] Black teachers who received a salary incentive in the 2009-10 school year increased their performance, on average, in the 2010-11 school year. This latter year is when community protests were strong. However, we remind readers that teachers made the decision about whether or not to opt into the salary incentive in the spring of the 2009-10 school year—in essence, already committing to improve their practice the next school year—whereas concerns and protests grew over the summer and into the fall.

first received an offer to opt into the salary incentive portion of the incentive system, called IMPACT*plus*. Opting in required giving up some job protections, and teachers could not reverse that decision in later years. On average, Black teachers opted into the salary incentive at substantially lower rates than White teachers (see Table 3). To the extent that opting in signals expectations of success, we hypothesize that we might see stronger incentive effects once accounting for this decision.

To disentangle the role of opting in from the initial offer, we apply a "fuzzy" regression discontinuity design framework where we use the immediate opt-in offer as an instrument for whether or not teachers actually did so in that school year. More specifically, we fit two-stage least squares models that simultaneously estimate one's probability of opting in based on the offer, and then use the predicted values of opting in to predict subsequent teacher performance. In Table 3, we show descriptive evidence that the opt-in offer and actually opting in are highly correlated, indicating a strong instrument.

The instrumental variables approach only makes sense to conduct in the first year of the evaluation system for several reasons. First, once teachers opted in, they could not reverse that decision in later years, even when they were eligible for a subsequent salary increase. Thus, it is most reasonable to focus on the first instance in which teachers received a salary incentive offer. Second, with the 2012-13 redesign, the introduction of a multi-tier career ladder provided teachers who earned an "Effective" rating (i.e., our comparison group) an opportunity to earn smaller salary increases that also required them to opt in and give up job protections. Because the opt-in offer was meant both for "Highly Effective" and "Effective" teachers, it would not make sense to use the "Highly Effective" rating as an instrument to predict opt-in offer acceptance. Third, estimation of a local average treatment effect for opt-in compliers is approximately the intent-to-treat estimate (shown above) divided by the take-up rate. Except for the first year of implementation, the intent-to-treat estimates for Black teachers are indistinguishable from zero. Therefore, we do not expect that estimation of a local average treatment effect through instrumental variables estimation would return any different results. Even if point estimates fluctuate to some extent when accounting for take-up, standard errors also will increase in a two-stage least squares analysis relative to the intent-to-treat analysis.

Results presented in Appendix Table 3 provide support for our hypothesis. In the first year of implementation, we often see larger effects of opting in on the subsequent performance of Black teachers (upwards of 0.62 SD for one of the least difficult tasks, Maximize Instructional Time) relative to White teachers (0.31 SD on this same task; p = 0.08 on test of difference of coefficients). When estimating salary incentive effects pooling across compliers and non-compliers, effects for Black teachers on this same task still are positive and statistically significant (0.29 SD; see Table 5). But, the point estimate is slightly smaller than the effect for White teachers (0.36 SD). In other words, for the relatively small subset of Black teachers who took up the opt-in offer in this first year (35%; see Table 3)—likely exhibiting greater expectations of success with the salary incentive—effects were quite large.

4.3. Robustness Tests

The validity of RD designs rests on several key assumptions. First, teachers cannot manipulate their evaluation score to move from one side of the eligibility threshold to another. Above, we provide evidence to suggest that this assumption holds in our data by comparing the observable characteristics of teachers and the schools they work in on either side of the thresholds that determine eligibility for the incentives. We expand this analysis in Appendix Table 4 by varying the sets of controls included in our models. Here and in additional robustness tests described below, we focus on a parsimonious set of models that pool data across all years. Compared to models that control only for year fixed effects and distance from the threshold, patterns of results are similar when we add observable teacher and school characteristics as controls (i.e., those characteristics listed in Table 1), as well as when we replace school characteristics with school fixed effects. We interpret these patterns as evidence that treatment and control groups are equal in expectation.

Second, researchers must correctly specify the functional form of the assignment variable in order to ensure that the "jump" in outcomes at the threshold is of the correct sign and magnitude. In our primary results, we specify a linear function of the forcing variable, supported in part by visual analysis of trends in Figure 2. In Appendix Table 4, we show that estimates are qualitatively similar when we instead specify a quadratic function of the forcing variable. We opt for a linear function of the forcing variable in our main analyses because the quadratic terms are not justified empirically. We more formally test the specification assumption in Appendix Tables 5 by reporting estimates that restrict the bandwidth to an increasingly narrow range around the performance thresholds. A narrower bandwidth decreases the model's reliance on functional form assumptions. We include results for bandwidths of 40, 30, and 20 points on either side of the threshold (our primary estimates come from a sample of teachers within 50 points on either side of the threshold). Our results do not always maintain statistical significance, given that the number of observations and statistical precision decrease. However, we see that the magnitudes of point estimates remain relatively stable, confirming that observations far from the performance threshold are not driving results.

One exception is that, at narrower bandwidths, we observe statistically significant negative effects of salary incentives on the performance of Black teachers on the task with the least expectations of success. While the magnitudes of these estimates increase in absolute value, the sign is the same across models. Our interpretation of patterns also is the same: in instances where Black teachers responded to the salary instances, we observe smaller effects on the tasks with the lowest expectations of success relative to tasks with higher expectations of success.

In a final robustness test, we re-estimate dismissal and salary incentive effects using classroom observation scores from lessons that school leaders (e.g., principals) rated versus lessons rated by a master educator from outside of the school and hired by the district specifically for the purpose of conducting these observations. This disaggregation probes the assumption that classroom observation scores reflect underlying teacher rather than rater behaviors. It is possible, for example, that school leaders who worked closely with teachers in their school may have been inclined to score lessons in way that achieved a desired result. This possibility is unlikely to be the case for master educators, who did not work closely with the teachers they observed, were less aware of their prior performance, and were unlikely to have a stake in whether a given teacher was dismissed or earned a salary increase.

In Appendix Table 6, we show that overall trends are quite similar when using as outcome measures observation scores generated from school leaders versus master educators. We exclude effects on the overall IMPACT score, as these scores cannot be disaggregated by rater. For dismissal threats, effect sizes are slightly larger when using scores from school leaders relative to master educators, though the differences are not statistically or substantively meaningful. When using scores from school leaders, we no longer detect differences in effect sizes for Black teachers on the least versus most difficult tasks. However, lack of statistically significant differences is explained, at least in part, by increased measurement error in observation scores calculated from fewer lessons, which limits precision. For salary incentives, there is some evidence of a reverse pattern: effect sizes are slightly smaller when using scores from school leaders compared to master educators. At the same time, results lead to the same substantive conclusions: salary incentives had large effects on the subsequent performance of White teachers (particularly on the most difficult tasks) but not for Black teachers, on average across school years.

5. Discussion and Conclusion

Consistent with the theoretical underpinnings of personnel economics (Lazear 2000; Holmstrom 1979), we find that job-embedded performance incentives can increase subsequent teacher performance in a way that is in the best interest of the employee (i.e., keeping their job, earning a higher salary) and the school (i.e., stronger classroom instruction and teacher quality for the thousands of students that DCPS serves). In some instances, the impacts may be considered quite large. We find dismissal threat and salary incentive effects on subsequent teacher performance as large as 0.6 SD, which is close to double the difference in performance between veteran and novice teachers in our dataset (0.35 SD). A more meaningful benchmark is the increase in teacher performance necessary to improve student outcomes. A growing body of evidence indicates that a SD increase in teaching quality results in a 0.1 to 0.2 SD increase in student test scores (Kane et al. 2011), and larger effects upwards of 0.3 SD on components of students' social-emotional development such as their engagement in class activities (Blazar and Kraft 2017). This implies that the effects on teaching tasks that we observe in our analyses likely are large enough to produce meaningful impacts on students. In the public sector field of education, student outcomes are a longstanding and common way to measure firm output (Hanushek 1979; Todd and Wolpin 2003).

However, the primary takeaway from this study should not be that performance incentives work, on average, but rather that there is substantial heterogeneity in their effects. We find heterogenous impacts along two dimensions—task difficulty and teacher race—which we argue are both related to employees' expectations of successfully achieving a desired goal. The fact that we often find larger effects of incentives on easier tasks with greatest expectation of success is intuitive and aligned to other lab-based literature on this topic (Locke and Latham 2002; Garber and Konradt 2014; Weibel, Rost, and Osterloh 2010). But, what implications does this have for schools as firms? From the traditional perspective of the principal-agent model, teachers' underinvestment in some tasks is seen as problem (Baker 1992; Baker, Gibbons, and Murphy 1994; Holmstrom and Milgrom 1991). Substituting effort away from the more difficult tasks and towards the easier ones means that teachers may not be attending to key parts of their jobs when faced with particularly strong incentives.

An alternative explanation for the differential effects of incentives on various tasks is that teachers exerted equal effort, but the more difficult tasks were substantially more difficult to improve relative to the easier tasks. While our study is not designed to distinguish between effort allocation and skill development, several pieces of data provide suggestive evidence for the former over the latter. First and foremost, differential effects on more versus less difficult tasks are not universal. For White teachers, dismissal threat effect sizes are indistinguishable across teaching tasks; we also find larger effects of the salary incentives on the more difficult tasks relative to the easier tasks. Second, for Black teachers for whom we do observe differential effects across task difficulty, patterns consistently point to a story related to expectancy and targeted effort in tasks and instances where expectations of success were highest. We return to discussion of social identity and its relation to expectancy below. Finally, in an analysis also drawing on the DCPS context and data, Phipps and Wiseman (2021) found that teachers worked to improve most during times of the school year when they were most likely to be observed.

Whether the differences in effects or dismissal threats and salary incentives across teaching tasks stems from targeted allocation of effort versus the innate nature of skill development should matter only in terms of overall output: in our case, the consequences for students. Like other professions, teaching is a multidimensional enterprise, and both theoretical and empirical work describes how students need teachers who excel in a range of teaching activities in order to support both short- and longer-term development (e.g., Jackson 2018). While more difficult tasks grounded in higher-level understanding may seem more desirable (simply because they are more difficult), in fact the literature suggests that both these and the less difficult tasks focused on supportive and efficient classroom environments benefit students. For example, Kane and colleagues (2011) identify effects on student test score performance in math and reading of roughly 0.1 SD for having a teacher who excels in building a supportive and efficient classroom environment. Supportive classroom environments also are central to theories on "culturally responsive" or "culturally relevant" teaching (Ladson-Billings 1995), and there is emerging experimental evidence that Black teachers are more likely to engage in these practices, which in turn drives race-matching effects on student outcomes (Blazar 2021). Other research identifies positive effects on student test scores in math and reading also around 0.1 SD for having a teacher who pushes for higher-level, conceptual understanding of academic content (Blazar 2015; Grossman et al. 2013). In other words, incentivizing improved teacher performance in multiple areas is valuable from the student and district perspective. While we cannot make concrete inferences beyond the teaching profession, we follow Holmstrom and Milgrom (1991) in hypothesizing that measurement of all desirable tasks in a pre-specified rubric may be the mechanism that leads to improvements in multiple job areas, even if improvements are not equal in magnitude across tasks.

More worrisome in terms of contract and incentive design are the differential effects we observe between Black and White teachers. Our findings are quite similar to a handful of other studies that show how incentives have less motivational value for individuals whose social identity has been subject to marginalization and discrimination (Farzana, Li, and Ren 2015; Hoff and Pandey 2006). A key finding from these lab-based experiments is that attenuated incentive effects for marginalized groups only emerges once their identity is revealed. In our setting, though—as in most on-the-job work—social identity markers such as race always are public. How then can school districts roll out incentives in a way that motivates all employees, and especially those from historically marginalized groups? Akerloff and Kranton (2000) theorized that social "outsiders" should receive larger monetary rewards than "insiders" in order to compensate them for acting in the interest of the firm rather than their own. However, designing contracts, pay scales, and incentives with differential compensation based on identity status is not practical because it is unlawful, at least in the U.S.[10]

That said, the ties between social identity and expectancy documented in this paper provide some potential guidance. *When expectations of success are high*, incentives can have very large effects on

[10] Civil Rights Act of 1964 § 7, 42 U.S.C. § 2000e et seq (1964).

the subsequent performance of Black teachers. Dismissal threats led to substantively and economically meaningful impacts of roughly 0.3 SD on tasks related to the classroom environment that have the greatest expectation of success. For salary incentives, we also observe very large impacts upwards of 0.6 SD—also on tasks related to the classroom environment—for Black teachers, at least amongst compliers who took-up the salary incentive opt-in offer and in the first year of implementation. Identifying how best to design incentive schemes to increase expectations of success for Black and other historically marginalized employees is beyond the immediate scope of this paper, but is an area ripe for additional analyses.

We conclude that expectancy is a driver of responses to motivation in on-the-job settings for teachers, and that social identity is one factor that can inform expectancy-based responses. To the extent that these results replicate for workers in multi-task jobs in other sectors, firms should consider adaptations to the traditional performance-based incentives model to account for social and racial identity and to expectancy.

References

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

Afridi, Farzana, Sherry Xin Li, and Yufei Ren. 2015. "Social Identity and Inequality: The Impact of China's Hukou System." *Journal of Public Economics* 123: 17-29.

Akerlof, George A., and Rachel E. Kranton. 2000. "Economics and Identity." *The Quarterly Journal of Economics* 115(3): 715-753.

----------. 2005. "Identity and the Economics of Organizations." *Journal of Economics Perspectives* 19(1): 9-32.

Atkinson, John W. 1957. "Motivational Determinants of Risk-Taking Behavior." *Psychological Review* 64(6p1): 359-372.

Baker, George P. 1992. "Incentive Contracts and Performance Measurement." *Journal of Political Economy* 100(3): 598-614.

Baker, George, Robert Gibbons, and Kevin J. Murphy. 1994. "Subjective Performance Measures in Optimal Incentive Contracts." *The Quarterly Journal of Economics* 109(4): 1125-1156.

Blazar, David. 2015. "Effective Teaching in Elementary Mathematics: Identifying Classroom Practices That Support Student Achievement." *Economics of Education Review* 48: 16-29.

Blazar, David. 2021. "Teachers of Color, Culturally Responsive Teaching, and Student Outcomes: Experimental Evidence from the Random Assignment of Teachers to Classes." Annenberg Institute at Brown University EdWorkingPaper 21-501.

Blazar, David, and Matthew A. Kraft. 2017. "Teacher and Teaching Effects on Students' Attitudes and Behaviors." *Educational Evaluation and Policy Analysis* 39(1): 146-170.

Bleiberg, Joshua, Eric Brunner, Erica Harbatkin, Matthew A. Kraft, and Matthew Springer. 2021.

"The Effect of Teacher Evaluation on Achievement and Attainment: Evidence from Statewide Reforms." Annenberg Institute at Brown University EdWorkingPaper 21-496.

Borjas, George. *Labor Economics*. Boston: McGraw Hill, 2020.

Boyd, Donald, Hamilton Lankford, Susanna Loeb, Jonah Rockoff, and James Wyckoff. 2008. "The narrowing gap in New York City teacher qualifications and its implications for student achievement in high-poverty schools." *Journal of Policy Analysis and Management,* 27: 793-818.

Bureau of Labor Statistics. 2020. "Industries with Largest Employment." United States Department of Labor.

Campbell, Donald T. 1969. "Reforms as Experiments." *American Psychologist* 24(4): 409-429.

Campbell, Shanyce L. Online 2020. "Ratings in Black and White: A Quantcrit Examination of Race and Gender in Teacher Evaluation Reform." *Race Ethnicity and Education*: 1-19.

Cardoza, Kavitha. "WTU Pres. Calls Teacher Evaluations 'Racist' Ahead of Ratings Release." *WAMU American University Radio*, June 30, 2011.

Danielson, Charlotte. *Enhancing Professional Practice: A Framework for Teaching.* Alexandria: Association for Supervision and Curriculum Development, 2011.

Dee, Thomas S. 2004. "Teachers, Race, and Student Achievement in a Randomized Experiment." *Review of Economics and Statistics* 86(1): 195-210.

Dee, Thomas S., and James Wyckoff. 2015. "Incentives, Selection, and Teacher Performance: Evidence from IMPACT." *Journal of Policy Analysis and Management* 34(2): 267-297.

Dee, Thomas S., Jessalynn James, and Jim Wyckoff. 2021. "Is Effective Teacher Evaluation Sustainable? Evidence from District of Columbia Public Schools." *Education Finance and Policy* 16(2): 313-346.

DeRamus-Byers, Raven. "Grow Your Own and Teacher Diversity in State Legislative Sessions: What We Can Learn from Successfully Passed Bills." *New America*, July 12, 2021.

District of Columbia Public Schools. "IMPACT Data Trends - Equity and Mitigating Implicit Bias." Accessed March 31, 2022.

District of Columbia Public Schools. "IMPACT Guidebook 2010-2011." Accessed March 31, 2022.

Education Commission of the States. "State Information Request: Diversifying the Teacher Workforce." October 21, 2019.

Ehrenberg, Ronald G., and Robert S. Smith. *Modern Labor Economics: Theory and Public Policy*. New York: Routledge, 2016.

Fryer Jr, Roland G. "The Production of Human Capital in Developed Countries: Evidence from 196 Randomized Field Experiments." In *Handbook of Economic Field Experiments*. Volume 2, edited by Banerjee, Abhijit V., and Esther Duflo, 95-322. Amsterdam: North-Holland, 2017.

Garbers, Yvonne, and Udo Konradt. 2014. "The Effect of Financial Incentives on Performance: A Quantitative Review of Individual and Team-Based Financial Incentives." *Journal of Occupational and Organizational Psychology* 87(1): 102-137.

Garrett, Rachel, and Matthew P. Steinberg. 2015. "Examining Teacher Effectiveness Using Classroom Observation Scores: Evidence from the Randomization of Teachers to Students." *Educational Evaluation and Policy Analysis* 37(2): 224-242.

Gay, Geneva. *Culturally Responsive Teaching: Theory, Research, and Practice*. New York: Teachers College Press, 2018.

Gneezy, Uri, Stephan Meier, and Pedro Rey-Biel. 2011. "When and Why Incentives (Don't) Work to Modify Behavior." *Journal of Economic Perspectives* 25(4): 191-210.

Grossman, Pam, Susanna Loeb, Julie Cohen, and James Wyckoff. 2013. "Measure for Measure: The Relationship between Measures of Instructional Practice in Middle School English Language Arts and Teachers' Value-Added Scores." *American Journal of Education* 119(3): 445-470.

Hamre, Bridget K., Robert C. Pianta, Jason T. Downer, Jamie DeCoster, Andrew J. Mashburn, Stephanie M. Jones, Joshua L. Brown, Elise Cappella, Marc Atkins, and Susan E. Rivers. 2013. "Teaching through Interactions: Testing a Developmental Framework of Teacher Effectiveness in over 4,000 Classrooms." *The Elementary School Journal* 113(4): 461-487.

Hanushek, Eric A. 1979. "Conceptual and Empirical Issues in the Estimation of Educational Production Functions." *Journal of Human Resources* 14(3): 351-388.

Hanushek, Eric A. 2007. "The Single Salary Schedule and Other Issues of Teacher Pay." *Peabody Journal of Education* 82(4): 574-586.

Hill, Heather C., Charalambos Y. Charalambous, and Matthew A. Kraft. 2012. "When Rater Reliability Is Not Enough: Teacher Observation Systems and a Case for the Generalizability Study." *Educational Researcher* 41(2): 56-64.

Hill, Heather C., David Blazar, and Kathleen Lynch. 2015. "Resources for Teaching: Examining Personal and Institutional Predictors of High-Quality Instruction." *AERA Open* 1(4): 2332858415617703.

Hoff, Karla, and Priyanka Pandey. 2006. "Discrimination, Social Identity, and Durable Inequalities." *American Economic Review* 96(2): 206-211.

Holmström, Bengt. 1979. "Moral Hazard and Observability." *The Bell Journal of Economics* 10(1): 74-91.

Holmstrom, Bengt, and Paul Milgrom. 1991. "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." *Journal of Law, Economics, and Organization* 7: 24.

Howell, William G., and Asya Magazinnik. 2017. "Presidential Prescriptions for State Policy: Obama's Race to the Top Initiative." *Journal of Policy Analysis and Management* 36(3): 502-531.

Imbens, Guido W., and Thomas Lemieux. 2008. "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics* 142(2): 615-635.

Jackson, C. Kirabo. 2018. "What Do Test Scores Miss? The Importance of Teacher Effects on Non–Test Score Outcomes." *Journal of Political Economy* 126(5): 2072-2107.

Jacob, Brian A., and Steven D. Levitt. 2003. "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *The Quarterly Journal of Economics* 118(3): 843-877.

Jiang, Jennie Y., and Susan E. Sporte. 2016. "Teacher evaluation in Chicago: Differences in observation and value-added scores by teacher, student, and school characteristics. Research Report." *University of Chicago Consortium on School Research*.

Kane, Thomas J., and Douglas O. Staiger. 2012. "Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Policy and Practice Brief. MET Project." *Bill & Melinda Gates Foundation*.

Kane, Thomas J., Eric S. Taylor, John H. Tyler, and Amy L. Wooten. 2011. "Identifying Effective Classroom Practices Using Student Achievement Data." *Journal of Human Resources* 46(3): 587-613.

Kaufman, Julia H., Benjamin K. Master, Alice Huguet, Paul Youngmin Yoo, Susannah Faxon-Mills, David Schulker, and Geoffrey E. Grimm. 2020. "Growing Teachers from Within: Implementation, Impact, and Cost of an Alternative Teacher Preparation Program in Three Urban School Districts." Santa Monica: RAND Corporation.

Knight, David S., and Thomas M. Skrtic. 2021. "Cost-Effectiveness of Instructional Coaching: Implementing a Design-Based, Continuous Improvement Model to Advance Teacher Professional Development." *Journal of School Leadership* 31(4): 318-342.

Kraft, Matthew A., David Blazar, and Dylan Hogan. 2018. "The Effect of Teacher Coaching on Instruction and Achievement: A Meta-Analysis of the Causal Evidence." *Review of Educational Research* 88(4): 547-588.

Kraft, Matthew A., and Allison F. Gilmour. 2017. "Revisiting the Widget Effect: Teacher Evaluation Reforms and the Distribution of Teacher Effectiveness." *Educational Researcher* 46(5): 234-249.

Ladson-Billings, Gloria. 1995. "Toward a Theory of Culturally Relevant Pedagogy." *American Educational Research Journal* 32(3): 465-491.

Lazear, Edward P. 2000. "The Power of Incentives." *American Economic Review* 90(2): 410-414.

----------. 2003. "Teacher Incentives." *Swedish Economic Policy Review* 10(2): 179-214.

Lee, David S., and Thomas Lemieux. 2010. "Regression Discontinuity Designs in Economics." *Journal of Economic Literature* 48(2): 281-355.

Levitt, Steven D., and John A. List. 2007. "What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?" *Journal of Economic Perspectives* 21(2): 153-174.

Lloyd, Robert, and Daniel Mertens. 2018. "Expecting More out of Expectancy Theory: History Urges Inclusion of the Social Context." *International Management Review* 14(1): 28-43.

Locke, Edwin A., and Gary P. Latham. 2002. "Building a Practically Useful Theory of Goal Setting and Task Motivation: A 35-Year Odyssey." *American Psychologist* 57(9): 705-717.

McCrary, Justin. 2008. "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test." *Journal of Econometrics* 142(2): 698-714.

McGuinn, Patrick. 2012. "Stimulating Reform: Race to the Top, Competitive Grants and the Obama Education Agenda." *Educational Policy* 26(1): 136-159.

Moore Johnson, Susan. 1984. "Merit Pay for Teachers: A Poor Prescription for Reform." *Harvard Educational Review* 54(2): 175-186.

Murname, Richard J., and D. Cohen. 1986. "Merit Pay and the Evaluation Problem: Understanding Why Most Merit Pay Plans Fail and a Few Survive." *Harvard Education Review* 56(1): 1-17.

Pham, Lam D., Tuan D. Nguyen, and Matthew G. Springer. 2021. "Teacher Merit Pay: A Meta-Analysis." *American Educational Research Journal* 58(3): 527-566.

Phipps, Aaron R., and Emily A. Wiseman. 2021. "Enacting the Rubric: Teacher Improvements in Windows of High-Stakes Observation." *Education Finance and Policy* 16(2): 283-312.

Pianta, Robert C., and Bridget K. Hamre. 2009. "Conceptualization, Measurement, and Improvement of Classroom Processes: Standardized Observation Can Leverage Capacity." *Educational Researcher* 38(2): 109-119.

Reardon, Sean F., and Joseph P. Robinson. 2012. "Regression Discontinuity Designs with Multiple Rating-Score Variables." *Journal of Research on Educational Effectiveness* 5(1): 83-104.

Sass, Tim R., Jane Hannaway, Zeyu Xu, David N. Figlio, and Li Feng. 2012. "Value Added of Teachers in High-Poverty Schools and Lower Poverty Schools." *Journal of Urban Economics* 72(2-3): 104-122.

Schaeffer, Katherine. "America's Public School Teachers are Far Less Racially and Ethnically Diverse than Their Students." *Pew Research Center*, December 10, 2021.

Shaw, Aaron D., John J. Horton, and Daniel L. Chen. 2011. "Designing Incentives for Inexpert Human Raters." In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, 275-284. New York: Association for Computing Machinery.

Steele, Claude M., Steven J. Spencer, and Joshua Aronson. 2002. "Contending with Group Image: The Psychology of Stereotype and Social Identity Threat." In *Advances in Experimental Social Psychology*, edited by Mark Zana, 379-441. Amsterdam: North-Holland.

Steinberg, Matthew P., and Lauren Sartain. 2021. "What Explains the Race Gap in Teacher Performance Ratings? Evidence from Chicago Public Schools." *Educational Evaluation and Policy Analysis* 43(1): 60-82.

Taylor, Eric S., and John H. Tyler. 2012. "The Effect of Evaluation on Teacher Performance." *American Economic Review* 102(7): 3628-51.

Todd, Petra E., and Kenneth I. Wolpin. 2003. "On the Specification and Estimation of the Production Function for Cognitive Achievement." *The Economic Journal* 113(485): F3-F33.

U.S. Department of Education. 2021. "Characteristics of Public School Teachers."

Weibel, Antoinette, Katja Rost, and Margit Osterloh. 2010. "Pay for Performance in the Public Sector—Benefits and (Hidden) Costs." *Journal of Public Administration Research and Theory* 20(2): 387-412.

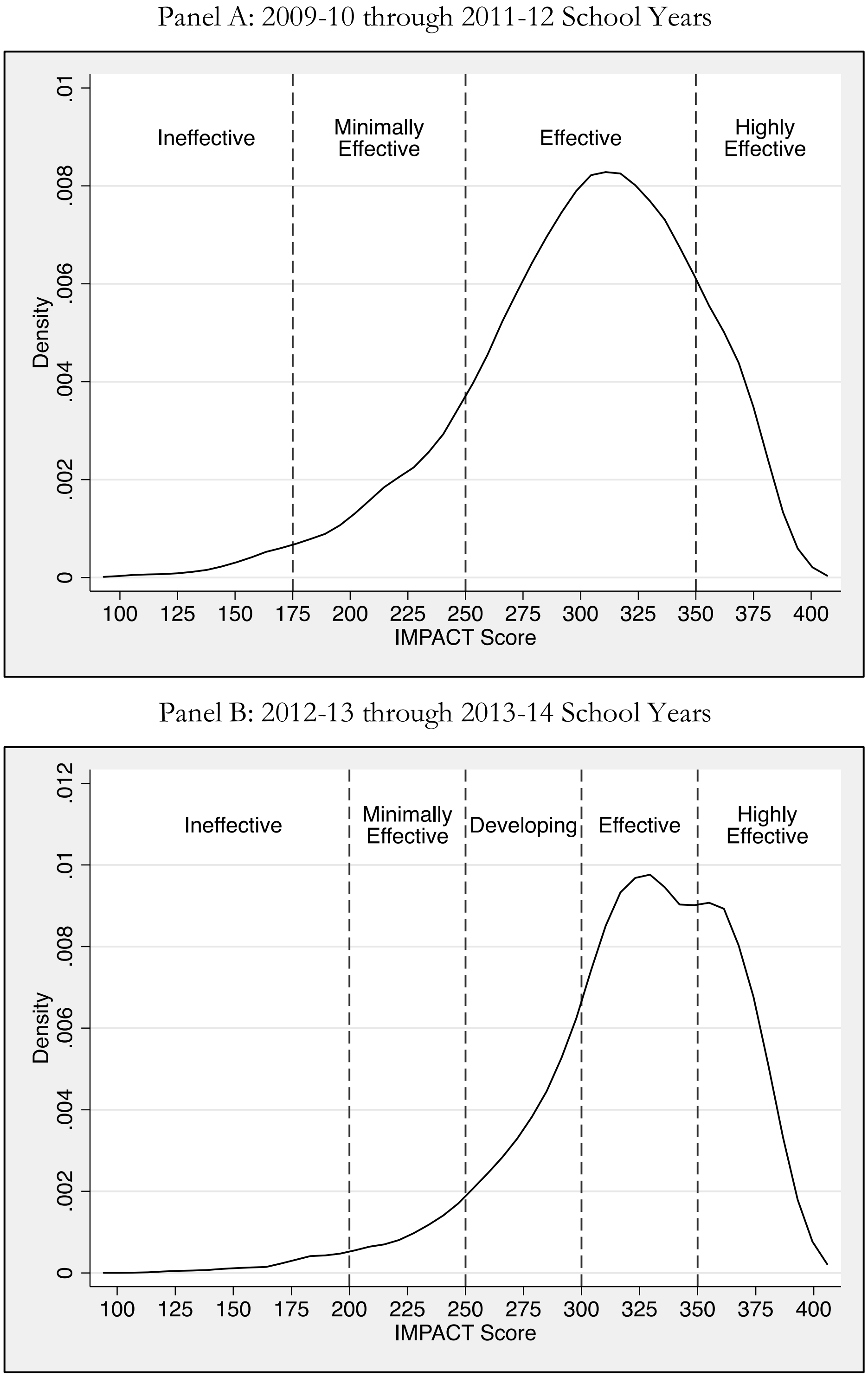
Weisberg, Daniel, Susan Sexton, Jennifer Mulhern, David Keeling, Joan Schunck, Ann Palcisco, and Kelli Morgan. 2009. "The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness." *New Teacher Project.*

Whitmire, Richard. *The Bee Eater: Michelle Rhee Takes on the Nation's Worst School District*. San Francisco: John Wiley & Sons, 2011.

Wiggins, Grant, and Jay McTighe. *Understanding by Design*. Alexandria: Association for Supervision and Curriculum Development, 2005.

Wong, Vivian C., Peter M. Steiner, and Thomas D. Cook. 2013. "Analyzing Regression-Discontinuity Designs with Multiple Assignment Variables: A Comparative Study of Four Estimation Methods." *Journal of Educational and Behavioral Statistics* 38(2): 107-141.

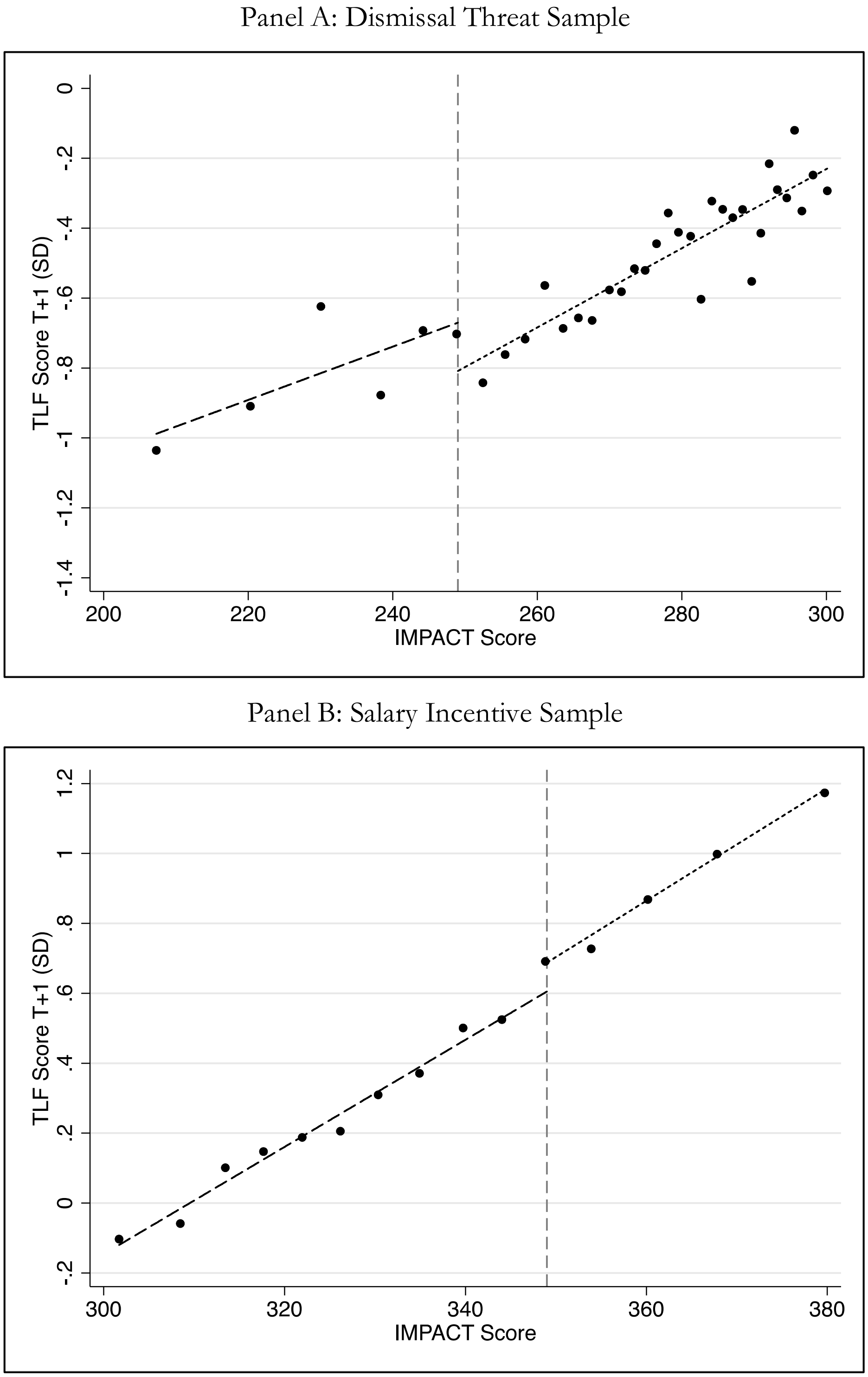
Figures



Panel A: 2009-10 through 2011-12 School Years

Panel B: 2012-13 through 2013-14 School Years

Figure 1: Distribution of summative IMPACT evaluation scores and associated ratings.



Panel A: Dismissal Threat Sample

Panel B: Salary Incentive Sample

Figure 2: RD impacts of dismissal threats (Panel A) and salary incentives (Panel B).

Tables

Table 1. Sample Descriptive Statistics

	All Teachers	Dismissal Threat Sample	Salary Incentive Sample
Female	0.74	0.70	0.77
Male	0.24	0.27	0.21
Gender Missing	0.03	0.03	0.02
Asian	0.03	0.03	0.04
Black	0.50	0.56	0.52
Hispanic	0.04	0.05	0.03
White	0.32	0.23	0.34
Race/Ethnicity Missing	0.11	0.14	0.07
Teaching Exp. 0 to 3 years	0.42	0.50	0.33
Teaching Exp.: 4 to 9 years	0.23	0.16	0.25
Teaching Exp. 10 years or more	0.34	0.32	0.41
Teaching Exp. Missing	0.01	0.01	0.01
Teach Gen. Ed., Tested Grades/Subjects	0.15	0.20	0.11
Teach Gen. Ed., Non-Tested Grades/Subjects	0.66	0.64	0.68
Teach Special Ed.	0.15	0.13	0.16
Teach English Language Learners	0.04	0.03	0.06
Low-Poverty School	0.77	0.90	0.76
High-Poverty School	0.23	0.10	0.24
Early Childhood	0.17	0.17	0.19
Elementary School	0.47	0.42	0.48
Middle School	0.26	0.27	0.25
High School	0.11	0.13	0.08
Dismissal Threat Sample	0.32	1.00	0.00
Dismissal Threat Offer	0.08	0.23	0.00
Salary Incentive Sample	0.33	0.00	1.00
Salary Incentive Offer	0.10	0.00	0.29
Has Outcome Data in Year T+1	0.79	0.69	0.85
Observations	19,469	4,885	6,418

Table 2. Descriptive Statistics on Practices from Teaching and Learning Framework

	Univariate Statistics						Intraclass Correlations (ICC)		
Tasks (Sorted from Lowest to Highest Score in Full Sample)	All Teachers		Dismissal Threat Sample		Salary Incentive Sample		Teacher-by-Year	Rater	Rater-by-Teacher Race
	Mean	SD	Mean	SD	Mean	SD			
Develop Higher-Level Understanding	2.72	0.59	2.25	0.46	2.96	0.43	0.636	0.160	0.037
Engage All Students in Rigorous Work	2.95	0.55	2.50	0.41	3.17	0.40	0.657	0.118	0.039
Respond to Misunderstandings	3.07	0.56	2.67	0.48	3.27	0.44	0.639	0.087	0.039
Explain Content Clearly	3.17	0.52	2.76	0.41	3.38	0.39	0.671	0.099	0.046
Lead Well-Organized Lessons	3.18	0.52	2.75	0.41	3.37	0.38	0.675	0.115	0.037
Provide Multiple Ways to Engage with Content	3.19	0.53	2.76	0.43	3.44	0.37	0.667	0.106	0.040
Check for Student Understanding	3.22	0.51	2.80	0.40	3.40	0.38	0.658	0.092	0.036
Maximize Instructional Time	3.27	0.57	2.80	0.48	3.50	0.39	0.698	0.084	0.028
Build a Supportive, Learning-Focused Classroom	3.40	0.51	2.99	0.46	3.57	0.35	0.703	0.110	0.029

Note: Following a generalizability framework, teacher-by-year ICC is adjusted for the modal number of lessons per teacher. The two least difficult and the two most difficult tasks (highlighted in gray) serve as outcome measures in subsequent analyses.

Table 3. IMPACT Consequences and Responses by Race and Year

School Year	Related to Dismissal		Related to Salary Increase	
	Immediately Separated	Threatened with Dismissal	Offered IMPACT*plus*	Opt-In if Offered
All Teachers				
All Years	0.03	0.08	0.10	0.66
2009	0.02	0.13	0.15	0.44
2010	0.06	0.12	0.05	0.77
2011	0.03	0.08	0.16	0.76
2012	0.02	0.06	0.05	0.62
2013	0.03	0.05	0.05	0.66
2014	0.04	0.04	0.10	0.70
Black Teachers				
All Years	0.03	0.08	0.08	0.58
2009	0.02	0.12	0.12	0.35
2010	0.05	0.13	0.04	0.67
2011	0.03	0.09	0.12	0.70
2012	0.03	0.07	0.06	0.51
2013	0.04	0.05	0.06	0.62
2014	0.04	0.04	0.10	0.67
White Teachers				
All Years	0.02	0.05	0.13	0.73
2009	0.01	0.11	0.22	0.55
2010	0.03	0.09	0.08	0.83
2011	0.02	0.05	0.24	0.83
2012	0.01	0.02	0.05	0.71
2013	0.01	0.03	0.06	0.72
2014	0.02	0.02	0.10	0.70

Note: Rows are shaded to identify time periods of evaluation system redesign. The evaluation system began in the 2009-10 school year. In the 2010-11 school year, small changes were made to the classroom observation rubric; no other changes were made in 2011-12 school year. In the 2012-13 school year, an additional performance band ("Developing") was added, splitting the middle part of the distribution ("Effective") into two groups (see Figure 1). In this same year, eligibility for salary increases was restricted to teachers working in schools where 60% or more of students were eligible for free or reduced-price meals; and weights for the different performance measures were adjusted to increase emphasis on classroom observations and decrease emphasis on teachers' contributions to student test-score growth.

Table 4. Regression Discontinuity (RD) Assumptions

	Dismissal Threat		Salary Incentive	
	Baseline	Balance	Baseline	Balance
	Balance	Yr. T+1	Balance	Yr. T+1
Female	0.007	0.007	0.001	0.003
	(0.005)	(0.006)	(0.005)	(0.006)
Gender Missing	-0.011	0.031	0.032*	0.032
	(0.021)	(0.051)	(0.016)	(0.028)
Asian	-0.005	0.014	-0.003	-0.007
	(0.012)	(0.015)	(0.011)	(0.012)
Black	0.005	0.003	-0.005	-0.004
	(0.006)	(0.006)	(0.006)	(0.006)
Hispanic	-0.002	0.004	0.018	0.023
	(0.011)	(0.013)	(0.013)	(0.014)
Race/Ethnicity Missing	0.013	0.006	0.001	-0.005
	(0.009)	(0.011)	(0.011)	(0.012)
Teaching Exp.: 0 to 3 years	-0.001	0.006	-0.000	-0.003
	(0.006)	(0.006)	(0.006)	(0.006)
Teaching Exp.: 4 to 9 years	0.001	-0.005	-0.005	-0.005
	(0.007)	(0.007)	(0.006)	(0.006)
Teaching Exp. Missing	0.010	-0.009	-0.011	0.081
	(0.029)	(0.058)	(0.024)	(0.080)
Teach Gen. Ed., Non-Tested Grades/Subj.	0.015*	0.009	-0.005	-0.002
	(0.006)	(0.007)	(0.007)	(0.008)
Teach Special Education	0.009	0.004	-0.010	-0.003
	(0.008)	(0.009)	(0.008)	(0.009)
Teach English Language Learners	0.007	-0.009	0.007	0.007
	(0.013)	(0.016)	(0.012)	(0.012)
High-Poverty School	-0.007	-0.014~	0.008	0.007
	(0.007)	(0.008)	(0.006)	(0.006)
Elementary School	-0.003	-0.008	-0.009	-0.009
	(0.006)	(0.007)	(0.006)	(0.007)
Middle School	-0.004	-0.009	-0.001	0.001
	(0.007)	(0.008)	(0.007)	(0.008)
High School	0.018*	0.013	0.000	0.001
	(0.008)	(0.010)	(0.009)	(0.010)
Observations	4,885	3,366	6,418	5,477
P-Value on Joint Test of Significance	0.186	0.176	0.182	0.601
P-Value on McCrary Test	0.934		0.456	

Notes: *** p<0.001, ** p<0.01, * p<0.05, ~ p<0.1. Estimates in each column are from separate regression models that predict a dummy indicator for dismissal threat or base salary increase offer as a function of the listed characteristics, year fixed effects, and distance from threshold. Heteroskedasticity-robust standard errors in parentheses. Left-out group categories include: male; White; teaching experience of 10 years or more; teach general education, non-tested grades and subjects; low-poverty school; and early childhood school.

Table 5. RD Estimates of Differential Responses to Evaluation Incentives by Task and Race

	Dismissal Threat				Salary Incentive			
Outcome Measures in Year T+1	All	Black	White	Black= White	All	Black	White	Black= White
Overall Scores								
IMPACT	0.206**	0.117	0.515***	0.000	0.069*	-0.010	0.226***	0.000
	(0.070)	(0.085)	(0.106)		(0.032)	(0.037)	(0.037)	
TLF	0.237***	0.215*	0.442***	0.040	0.077*	0.009	0.206***	0.000
	(0.071)	(0.084)	(0.114)		(0.034)	(0.039)	(0.040)	
Most Difficult Tasks								
Higher-Level Understanding (HLU)	0.164*	0.124	0.447***	0.004	0.099*	-0.015	0.296***	0.000
	(0.071)	(0.084)	(0.113)		(0.040)	(0.046)	(0.049)	
Rigorous Work (RW)	0.145*	0.154~	0.284*	0.255	0.053	-0.005	0.182***	0.000
	(0.071)	(0.081)	(0.121)		(0.040)	(0.046)	(0.047)	
Least Difficult Tasks								
Maximize Instructional Time (MIT)	0.297***	0.285**	0.361**	0.510	0.040	0.040	0.075~	0.342
	(0.077)	(0.092)	(0.120)		(0.038)	(0.042)	(0.045)	
Supportive Classroom (SC)	0.292***	0.326***	0.389**	0.605	0.017	0.024	0.033	0.791
	(0.079)	(0.095)	(0.122)		(0.037)	(0.041)	(0.045)	
P-Value on Tests Between Practices:								
HLU=MIT	0.047	0.040	0.405		0.160	0.243	0.000	
HLU=SC	0.083	0.018	0.613		0.054	0.419	0.000	
RW=MIT	0.012	0.061	0.409		0.734	0.327	0.024	
RW=SC	0.036	0.032	0.378		0.389	0.562	0.004	
Observations	3,366	2,037	754		5,477	2,995	1,822	

Notes: *** p<0.001, ** p<0.01, * p<0.05, ~ p<0.1. For each incentive and teacher performance measure, estimates for sample of all teachers come from separate regression models that control for distance from the threshold; estimates for Black versus White teachers come from the same regression models that control for distance from the threshold interacted with teacher race dummies. All models control for year fixed effects. Estimates are standardized effect sizes. Heteroskedasticity-robust standard errors in parentheses.

Table 6. RD Estimates of Differential Responses to Salary Incentives by Task, Race, and Time Period

Outcome Measures in Year T+1	2009-10				2010-11 and 2011-12				2012-13 through 2014-15			
	All	Black	White	Black= White	All	Black	White	Black= White	All	Black	White	Black= White
Overall Scores												
IMPACT	0.097~	0.045	0.251***	0.001	0.050	-0.061	0.216***	0.000	0.127*	0.064	0.256**	0.003
	(0.059)	(0.072)	(0.065)		(0.048)	(0.058)	(0.054)		(0.062)	(0.068)	(0.084)	
TLF	0.102~	0.054	0.232***	0.002	0.059	-0.015	0.193***	0.000	0.171*	0.100	0.273**	0.018
	(0.060)	(0.070)	(0.067)		(0.050)	(0.059)	(0.057)		(0.070)	(0.076)	(0.094)	
Most Difficult Tasks												
Higher-Level Understanding (HLU)	0.095	-0.038	0.286***	0.000	0.129*	0.021	0.319***	0.000	0.124	0.042	0.305**	0.003
	(0.074)	(0.089)	(0.085)		(0.059)	(0.071)	(0.070)		(0.082)	(0.090)	(0.112)	
Rigorous Work (RW)	0.075	0.083	0.132	0.503	0.040	-0.055	0.220**	0.000	0.125	0.061	0.232*	0.039
	(0.072)	(0.086)	(0.084)		(0.060)	(0.071)	(0.069)		(0.080)	(0.087)	(0.104)	
Least Difficult Tasks												
Maximize Instructional Time (MIT)	0.154*	0.229**	0.190*	0.545	-0.023	-0.078	0.045	0.033	0.100	0.099	0.035	0.398
	(0.068)	(0.076)	(0.081)		(0.056)	(0.066)	(0.066)		0.054	(0.082)	(0.100)	
Supportive Classroom (SC)	0.146*	0.158*	0.209**	0.405	-0.061	-0.036	-0.047	0.840	0.084	0.047	0.030	0.822
	(0.067)	(0.078)	(0.077)		(0.055)	(0.064)	(0.066)		(0.073)	(0.079)	(0.098)	
P-Value on Tests Between Practices:												
HLU=MIT	0.451	0.002	0.311		0.014	0.187	0.000		0.774	0.540	0.009	
HLU=SC	0.521	0.031	0.401		0.002	0.439	0.000		0.640	0.960	0.010	
RW=MIT	0.245	0.074	0.477		0.280	0.753	0.014		0.760	0.686	0.055	
RW=SC	0.337	0.394	0.383		0.117	0.807	0.001		0.620	0.881	0.055	
Observations	1,446	757	521		2,712	1,412	955		1,319	826	346	

Notes: *** $p<0.001$, ** $p<0.01$, * $p<0.05$, ~ $p<0.1$. For each teacher performance measure and time period, estimates for sample of all teachers come from separate regression models that control for distance from the threshold; estimates for Black versus White teachers come from the same regression models that control for distance from the threshold interacted with teacher race dummies. In models with multiple school years, year fixed effects are included. Estimates are standardized effect sizes. Heteroskedasticity-robust standard errors in parentheses.

Appendix

Appendix Table 1. Distribution of Teachers between High- versus Low-Poverty Schools

	High-Poverty Schools	Low-Poverty Schools
Teacher Race		
Asian	0.03	0.04
Black	0.56	0.30
Hispanic	0.04	0.04
White	0.25	0.54
Race/Ethnicity Missing	0.11	0.08
IMPACT Rating		
Ineffective	0.02	0.01
Minimally Effective	0.11	0.03
Developing	0.11	0.04
Effective	0.59	0.48
Highly Effective	0.17	0.44
Observations	14,930	4,480

Note: School poverty rate is defined by DCPS administration as the percent of students eligible for free or reduced-price meals, with a cutoff at 60%.

Appendix Table 2. RD Estimates of Differential Responses to Dismissal Threats by Task, Race, and Time Period

	2010-11 and 2011-12				2012-13 through 2014-15			
Outcome Measures in Year T+1	All	Black	White	Black= White	All	Black	White	Black= White
Overall Scores								
IMPACT	0.201*	0.107	0.531***	0.001	0.199~	0.123	0.441*	0.078
	(0.089)	(0.105)	(0.130)		(0.113)	(0.140)	(0.181)	
TLF	0.246**	0.218*	0.515***	0.033	0.212~	0.211	0.270	0.739
	(0.091)	(0.105)	(0.142)		(0.112)	(0.137)	(0.183)	
Most Difficult Tasks								
Higher-Level Understanding (HLU)	0.198*	0.114	0.595***	0.001	0.105	0.149	0.152	0.985
	(0.093)	(0.108)	(0.143)		(0.108)	(0.133)	(0.178)	
Rigorous Work (RW)	0.178~	0.174~	0.373*	0.174	0.088	0.123	0.101	0.903
	(0.095)	(0.106)	(0.155)		(0.108)	(0.126)	(0.191)	
Least Difficult Tasks								
Maximize Instructional Time (MIT)	0.246*	0.242*	0.378**	0.330	0.355**	0.340*	0.274	0.744
	(0.098)	(0.112)	(0.146)		(0.125)	(0.155)	(0.209)	
Supportive Classroom (SC)	0.286**	0.343**	0.429**	0.559	0.288*	0.308*	0.242	0.749
	(0.098)	(0.119)	(0.147)		(0.130)	(0.156)	(0.210)	
P-Value on Tests Between Practices:								
HLU=MIT	0.574	0.205	0.090		0.019	0.123	0.485	
HLU=SC	0.365	0.044	0.265		0.107	0.220	0.614	
RW=MIT	0.384	0.452	0.966		0.005	0.051	0.244	
RW=SC	0.247	0.119	0.724		0.058	0.123	0.420	
Observations	1,826	1,095	455		1,540	942	299	

Notes: *** p<0.001, ** p<0.01, * p<0.05, ~ p<0.1. For each teacher performance measure and time period, estimates for sample of all teachers come from separate regression models that control for distance from the threshold; estimates for Black versus White teachers come from the same regression models that control for distance from the threshold interacted with teacher race dummies. In models with multiple school years, year fixed effects are included. Estimates are standardized effect sizes. Heteroskedasticity-robust standard errors in parentheses.

Appendix Table 3. IV Estimates of Effect of Opting into Salary Incentive (2009-10 School Year)

Outcome Measures in Year T+1	All	Black	White	Black= White
Overall Scores				
IMPACT	0.191	0.092	0.401***	0.052
	(0.117)	(0.198)	(0.108)	
TLF	0.202~	0.123	0.371***	0.102
	(0.119)	(0.194)	(0.111)	
Most Difficult Tasks				
Higher-Level Understanding (HLU)	0.189	0.148	0.453**	0.003
	(0.147)	(0.245)	(0.140)	
Rigorous Work (RW)	0.148	0.218	0.215	0.986
	(0.142)	(0.240)	(0.136)	
Least Difficult Tasks				
Maximize Instructional Time (MIT)	0.305*	0.623**	0.313*	0.077
	(0.137)	(0.226)	(0.133)	
Supportive Classroom (SC)	0.288*	0.420~	0.340**	0.637
	(0.135)	(0.222)	(0.127)	
Observations	1,446	757	521	

Notes: *** p<0.001, ** p<0.01, * p<0.05, ~ p<0.1. IV estimates use exogenous variation in performance band right around the threshold to instrument for teachers' decision of whether or not to opt into the salary incentive portion of the evaluation system. For each teacher performance measure, estimates for sample of all teachers come from separate regression models that control for distance from the threshold; estimates for Black versus White teachers come from the same regression models that control for distance from the threshold interacted with teacher race dummies. All estimates are standardized effect sizes. Heteroskedasticity-robust standard errors in parentheses.

Appendix Table 4. Sensitivity of RD Estimates to Control Set and Functional Form

Outcome Measures in Year T+1	Observable Teacher and School Characteristics				Observable Teacher Characteristics and School Fixed Effects				Quadratic Function of Forcing Variable			
	All	Black	White	Black= White	All	Black	White	Black= White	All	Black	White	Black= White
						Dismissal Threat						
Overall Scores												
IMPACT	0.202**	0.138	0.473***	0.001	0.169*	0.132	0.422***	0.006	0.216**	0.161	0.561***	0.000
	(0.070)	(0.084)	(0.105)		(0.069)	(0.085)	(0.109)		(0.083)	(0.100)	(0.119)	
TLF	0.227**	0.237**	0.473***	0.169	0.208**	0.241**	0.355**	0.300	0.210*	0.225*	0.453***	0.039
	(0.070)	(0.082)	0.473***		(0.070)	(0.083)	(0.115)		(0.083)	(0.097)	(0.124)	
Most Difficult Tasks												
Higher-Level Understanding (HLU)	0.158*	0.137~	0.414***	0.010	0.131~	0.124	0.371**	0.027	0.149~	0.142	0.465***	0.004
	(0.070)	(0.083)	(0.111)		(0.070)	(0.083)	(0.114)		(0.082)	(0.097)	(0.125)	
Rigorous Work (RW)	0.158*	0.183*	0.218~	0.759	0.131~	0.199*	0.202~	0.983	0.136~	0.165~	0.295*	0.254
	0.158*	(0.080)	(0.120)		(0.071)	(0.081)	(0.122)		(0.081)	(0.091)	(0.131)	
Least Difficult Tasks												
Maximize Instructional Time (MIT)	0.289***	0.302***	0.328**	0.824	0.269***	0.287**	0.323**	0.755	0.256**	0.273**	0.349**	0.513
	(0.077)	(0.090)	(0.121)		(0.077)	(0.091)	(0.122)		(0.089)	(0.104)	(0.131)	
Supportive Classroom (SC)	0.292***	0.343***	0.376**	0.781	0.273***	0.329***	0.349**	0.871	0.254**	0.305**	0.367**	0.610
	(0.078)	(0.094)	(0.119)		(0.078)	(0.094)	(0.122)		(0.091)	(0.109)	(0.132)	
P-Value on Tests Between Practices:												
HLU=MIT	0.050	0.034	0.410		0.035	0.035	0.647		0.166	0.138	0.294	
HLU=SC	0.069	0.015	0.736		0.048	0.014	0.846		0.222	0.095	0.440	
RW=MIT	0.013	0.085	0.246		0.020	0.199	0.187		0.078	0.153	0.595	
RW=SC	0.028	0.045	0.179		0.037	0.100	0.191		0.141	0.124	0.588	
Observations	3,366	2,037	754		3,366	2,037	754		3,366	2,037	754	

						Salary Incentive						
Overall Scores												
IMPACT	0.065*	0.011	0.189***	0.000	0.053~	0.048	0.116**	0.060	0.078*	-0.017	0.219***	0.000
	(0.032)	(0.037)	(0.038)		(0.032)	(0.037)	(0.038)		(0.032)	(0.038)	(0.038)	
TLF	0.068*	0.036	0.150***	0.002	0.049	0.048	0.099*	0.190	0.089*	0.005	0.201***	0.000
	(0.034)	(0.039)	(0.041)		(0.034)	(0.039)	(0.042)		(0.035)	(0.040)	(0.041)	
Most Difficult Tasks												
Higher-Level Understanding (HLU)	0.085*	0.014	0.225***	0.000	0.074~	0.047	0.154**	0.030	0.110*	-0.016	0.295***	0.000
	(0.040)	(0.046)	(0.050)		(0.040)	(0.047)	(0.051)		(0.043)	(0.049)	(0.052)	
Rigorous Work (RW)	0.049	0.035	0.115*	0.076	0.039	0.051	0.080	0.547	0.054	-0.023	0.163**	0.000
	(0.039)	(0.046)	(0.049)		(0.040)	(0.047)	(0.050)		(0.042)	(0.048)	(0.050)	
Least Difficult Tasks												
Maximize Instructional Time (MIT)	0.036	0.065	0.029	0.366	0.011	0.052	-0.005	0.197	0.076*	0.063	0.098*	0.333
	(0.038)	(0.042)	(0.046)		(0.037)	(0.043)	(0.048)		0.054	(0.042)	(0.046)	
Supportive Classroom (SC)	0.016	0.035	0.017	0.646	-0.007	0.034	-0.034	0.109	0.049	0.041	0.051	0.780
	(0.037)	(0.042)	(0.046)		(0.037)	(0.042)	(0.046)		(0.037)	(0.042)	(0.045)	
P-Value on Tests Between Practices:												
HLU=MIT	0.240	0.283	0.000		0.122	0.923	0.003		0.445	0.122	0.000	
HLU=SC	0.099	0.670	0.000		0.049	0.795	0.000		0.168	0.272	0.000	
RW=MIT	0.744	0.519	0.079		0.467	0.989	0.084		0.609	0.078	0.193	
RW=SC	0.428	0.986	0.062		0.261	0.733	0.033		0.897	0.226	0.039	
Observations	5,477	2,995	1,822		5,477	2,995	1,822		5,477	2,995	1,822	

Notes: *** p<0.001, ** p<0.01, * p<0.05, ~ p<0.1. For each incentive and teacher performance measure, estimates for sample of all teachers come from separate regression models that control for distance from the threshold; estimates for Black versus White teachers come from the same regression models that control for distance from the threshold interacted with teacher race dummies. All models control for year fixed effects. Additional sets of covariates are listed in the column headers; the specific teacher and school characteristics are those listed in Table 5. Estimates are standardized effect sizes. Heteroskedasticity-robust standard errors in parentheses.

Appendix Table 5. Sensitivity of RD Estimates to Bandwidth

Outcome Measures in Year T+1	+/- 40				+/- 30				+/- 20			
	All	Black	White	Black= White	All	Black	White	Black= White	All	Black	White	Black= White
Dismissal Threat												
Overall Scores												
IMPACT	0.202*	0.119	0.537***	0.000	0.254**	0.179	0.596***	0.001	0.217~	0.175	0.517**	0.014
	(0.081)	(0.096)	(0.117)		(0.094)	(0.110)	(0.133)		(0.116)	(0.136)	(0.161)	
TLF	0.246**	0.241*	0.473***	0.150	0.281**	0.269*	0.499***	0.064	0.296*	0.276*	0.461**	0.196
	(0.082)	(0.096)	0.473***		(0.095)	(0.110)	(0.139)		(0.119)	(0.137)	(0.169)	
Most Difficult Tasks												
Higher-Level Understanding (HLU)	0.182*	0.155~	0.459***	0.010	0.273**	0.229*	0.593***	0.005	0.213~	0.135	0.493**	0.018
	(0.080)	(0.094)	(0.124)		(0.093)	(0.107)	(0.141)		(0.115)	(0.133)	(0.171)	
Rigorous Work (RW)	0.158*	0.222*	0.280*	0.629	0.186*	0.221*	0.296~	0.575	0.180	0.193	0.305	0.469
	0.158*	(0.092)	(0.133)		(0.094)	(0.105)	(0.152)		(0.117)	(0.130)	(0.187)	
Least Difficult Tasks												
Maximize Instructional Time (MIT)	0.339***	0.347***	0.371**	0.845	0.360***	0.396**	0.421**	0.858	0.395**	0.458**	0.393*	0.669
	(0.089)	(0.105)	(0.133)		(0.103)	(0.122)	(0.149)		(0.126)	(0.150)	(0.176)	
Supportive Classroom (SC)	0.317***	0.365***	0.353**	0.928	0.323**	0.377**	0.436**	0.668	0.405**	0.509**	0.437*	0.632
	(0.090)	(0.108)	(0.133)		(0.106)	(0.126)	(0.149)		(0.130)	(0.154)	(0.176)	
P-Value on Tests Between Practices:												
HLU=MIT	0.036	0.027	0.440		0.313	0.094	0.178		0.083	0.007	0.514	
HLU=SC	0.107	0.029	0.405		0.608	0.183	0.280		0.112	0.006	0.751	
RW=MIT	0.044	0.115	0.385		0.031	0.058	0.307		0.033	0.021	0.558	
RW=SC	0.141	0.122	0.581		0.138	0.141	0.351		0.051	0.015	0.472	
Observations	2,458	1,498	532		1,733	1,048	365		1,078	648	218	

						Salary Incentive						
Overall Scores												
IMPACT	0.082*	-0.012	0.231***	0.000	0.059	-0.042	0.226***	0.000	0.039	-0.080	0.211***	0.000
	(0.034)	(0.039)	(0.040)		(0.038)	(0.044)	(0.046)		(0.046)	(0.052)	(0.056)	
TLF	0.083*	-0.002	0.197***	0.000	0.054	-0.044	0.169***	0.000	0.019	-0.083	0.121*	0.000
	(0.036)	(0.041)	(0.043)		(0.041)	(0.046)	(0.049)		(0.049)	(0.055)	(0.060)	
Most Difficult Tasks												
Higher-Level Understanding (HLU)	0.097*	-0.040	0.294***	0.000	0.061	-0.093~	0.284***	0.000	-0.018	-0.194**	0.187**	0.000
	(0.044)	(0.050)	(0.053)		(0.050)	(0.056)	(0.060)		(0.060)	(0.067)	(0.072)	
Rigorous Work (RW)	0.038	-0.036	0.153**	0.000	0.026	-0.065	0.156**	0.000	0.021	-0.090	0.143*	0.000
	(0.043)	(0.050)	(0.052)		(0.049)	(0.056)	(0.058)		(0.059)	(0.066)	(0.071)	
Least Difficult Tasks												
Maximize Instructional Time (MIT)	0.068~	0.055	0.091~	0.358	0.057	0.035	0.072	0.401	0.064	0.046	0.049	0.957
	(0.041)	(0.045)	(0.049)		(0.046)	(0.050)	(0.056)		0.054	(0.060)	(0.068)	
Supportive Classroom (SC)	0.046	0.046	0.047	0.970	0.039	0.038	0.036	0.964	-0.023	-0.012	-0.043	0.568
	(0.040)	(0.044)	(0.048)		(0.045)	(0.050)	(0.054)		(0.054)	(0.059)	(0.067)	
P-Value on Tests Between Practices:												
HLU=MIT	0.535	0.067	0.000		0.926	0.026	0.001		0.188	0.000	0.071	
HLU=SC	0.269	0.103	0.000		0.670	0.028	0.000		0.935	0.010	0.002	
RW=MIT	0.486	0.069	0.233		0.535	0.071	0.149		0.461	0.042	0.189	
RW=SC	0.865	0.129	0.060		0.810	0.092	0.056		0.490	0.287	0.016	
Observations	4,705	2,543	1,596		3,790	2,024	1,314		2,624	1,384	932	

Notes: *** p<0.001, ** p<0.01, * p<0.05, ~ p<0.1. For each incentive and teacher performance measure, estimates for sample of all teachers come from separate regression models that control for distance from the threshold; estimates for Black versus White teachers come from the same regression models that control for distance from the threshold interacted with teacher race dummies. All models control for year fixed effects. Estimates are standardized effect sizes. Heteroskedasticity-robust standard errors in parentheses.

Appendix Table 6. RD Estimates of Differential Responses to Evaluation Incentives by Race, Task, and Rater Type

Outcome Measures in Year T+1	Rater is School Leader				Rater is Master Educator			
	All	Black	White	Black= White	All	Black	White	Black= White
Dismissal Threat								
Overall Scores								
TLF	0.254**	0.179~	0.435**	0.056	0.183**	0.194**	0.349***	0.000
	(0.087)	(0.106)	(0.135)		(0.063)	(0.073)	(0.101)	
Most Difficult Tasks								
Higher-Level Understanding (HLU)	0.187*	0.137	0.450***	0.017	0.092	0.074	0.302**	0.018
	(0.085)	(0.102)	(0.134)		(0.062)	(0.074)	(0.100)	
Rigorous Work (RW)	0.165~	0.184~	0.222	0.783	0.086	0.080	0.251*	0.065
	(0.087)	(0.102)	(0.143)		(0.062)	(0.071)	(0.098)	
Least Difficult Tasks								
Maximize Instructional Time (MIT)	0.349***	0.273*	0.446**	0.202	0.173*	0.217**	0.182~	0.725
	(0.090)	(0.109)	(0.139)		(0.067)	(0.080)	(0.106)	
Supportive Classroom (SC)	0.340***	0.362**	0.388**	0.859	0.164*	0.197*	0.256*	0.551
	(0.096)	(0.117)	(0.149)		(0.064)	(0.077)	(0.102)	
P-Value on Tests Between Practices:								
HLU=MIT	0.048	0.159	0.973		0.232	0.068	0.267	
HLU=SC	0.101	0.039	0.680		0.302	0.128	0.676	
RW=MIT	0.018	0.323	0.058		0.155	0.055	0.467	
RW=SC	0.048	0.085	0.261		0.239	0.126	0.962	
Observations	3,353	2,038	752		3,359	2,031	754	

	Salary Incentive							
Overall Scores								
TLF	0.042	0.014	0.100*	0.022	0.078*	-0.007	0.234***	0.000
	(0.038)	(0.043)	(0.046)		(0.033)	(0.037)	(0.038)	
Most Difficult Tasks								
Higher-Level Understanding (HLU)	0.052	-0.014	0.195***	0.000	0.101**	-0.022	0.282***	0.000
	(0.045)	(0.052)	(0.055)		(0.039)	(0.045)	(0.047)	
Rigorous Work (RW)	0.025	-0.002	0.069	0.148	0.056	-0.008	0.218***	0.000
	(0.047)	(0.054)	(0.057)		(0.036)	(0.042)	(0.044)	
Least Difficult Tasks								
Maximize Instructional Time (MIT)	0.019	0.010	0.043	0.438	0.032	0.043	0.066	0.487
	(0.043)	(0.049)	(0.052)		(0.035)	(0.039)	(0.041)	
Supportive Classroom (SC)	-0.031	-0.033	-0.034	0.989	0.046	0.057	0.077~	0.550
	(0.043)	(0.049)	(0.052)		(0.033)	(0.038)	(0.040)	
P-Value on Tests Between Practices:								
HLU=MIT	0.519	0.684	0.015		0.101	0.178	0.000	
HLU=SC	0.109	0.743	0.000		0.183	0.101	0.000	
RW=MIT	0.905	0.832	0.670		0.525	0.237	0.001	
RW=SC	0.293	0.613	0.112		0.794	0.153	0.003	
Observations	5,468	2,990	1,819		5,455	2,979	1,819	

Notes: *** p<0.001, ** p<0.01, * p<0.05, ~ p<0.1. For each incentive, teacher performance measure, and time period, estimates for sample of all teachers come from separate regression models that control for distance from the threshold; estimates for Black versus White teachers come from the same regression models that control for distance from the threshold interacted with teacher race dummies. All models control for year fixed effects. Estimates are standardized effect sizes. Heteroskedasticity-robust standard errors in parentheses.