Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 24-1068

Differential Responses to Teacher Evaluation Incentives: Expectancy, Race, Experience, and Task

David Blazar
University of Maryland, College Park

Melinda Adnot
University of North Carolina, Charlotte

Max Anthenelli
University of Maryland, College Park

Xinyi Zhong
University of Washington, Seattle

Teacher evaluation systems and their associated incentives have produced fairly mixed results. Our analyses are motivated by theory and descriptive evidence that accountability systems are highly racialized, and that individuals are less likely to respond to incentives when they have low expectations of success (and vice versa). Using a regression discontinuity design, we find that Black novices in the District of Columbia Public Schools faced the most negative consequences (dismissal threats) and the least benefits (salary incentives), without responding to either. White novices, in contrast, exhibited high expectations of success and large behavior changes, particularly in response to dismissal threats (0.6 SD). We also find some evidence of heterogeneity in effects by task difficulty, though these differences are less stark.

VERSION: October 2024

Suggested citation: Blazar, David, Melinda Adnot, Max Anthenelli, and Xinyi Zhong. (2024). Differential Responses to Teacher Evaluation Incentives: Expectancy, Race, Experience, and Task. (EdWorkingPaper: 24-1068). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/sr1d-7s24

Differential Responses to Teacher Evaluation Incentives: Expectancy, Race, Experience, and Task

David Blazar[1], Melinda Adnot[2], Max Anthenelli[1], and Xinyi Zhong[3]

Abstract

Teacher evaluation systems and their associated incentives have produced fairly mixed results. Our analyses are motivated by theory and descriptive evidence that accountability systems are highly racialized, and that individuals are less likely to respond to incentives when they have low expectations of success (and vice versa). Using a regression discontinuity design, we find that Black novices in the District of Columbia Public Schools faced the most negative consequences (dismissal threats) and the least benefits (salary incentives), without responding to either. White novices, in contrast, exhibited high expectations of success and large behavior changes, particularly in response to dismissal threats (0.6 SD). We also find some evidence of heterogeneity in effects by task difficulty, though these differences are less stark.

Keywords: teacher evaluation, incentives, race, expectancy theory

[1]University of Maryland, College Park; [2]University of North Carolina, Charlotte; [3]University of Washington, Seattle.
The research presented in this paper 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.

Introduction

Over the past 15 years, U.S. school districts—as well as state agencies and the federal government—have invested heavily in teacher evaluation and incentive schemes (Howell & Magazinnik, 2017; McGuinn, 2012) with very mixed results. Early and small-scale pilot programs demonstrated some successes at improving teacher performance and student outcomes (e.g., Steinberg & Sartain, 2015; Taylor & Tyler, 2012). Yet, the scale-up of teacher evaluation reforms across states has produced null effects on average district performance (Bleiberg et al., 2023). One explanation for this pattern is low fidelity of implementation: almost all teachers receive satisfactory ratings and there is very little action (e.g., dismissal) in response to low performance (Kraft & Gilmour, 2017). Studies that focus more narrowly on the merit pay component of teacher evaluation systems for high-performers tend to find more encouraging results (for a meta-analysis, see 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 student-level standard deviations [SD]) relative to other performance-enhancing interventions for teachers that are similar in cost (upwards of 0.2 student-level SD and 0.5 teacher-level SD on teacher performance measures; Fryer, 2017; Kraft, Blazar, & Hogan, 2018).

We add to the literature on teacher evaluation successes and failures by examining heterogeneity in effects. Whereas other studies examine heterogeneity by programmatic design features (e.g., adding a professional development component, magnitude of the financial incentive; Bleiberg et al., 2023; Pham et al, 2021), we focus on characteristics of individual teachers and the teaching tasks for which they are held accountable. Drawing on conceptual frameworks related to school-based accountability and equity (e.g., Darling-Hammond, 2007), expectancy theory and the psychological determinants of risk-taking behavior (e.g., Atkinson, 1957; Locke & Latham, 2002),

and personnel and identity economics (e.g., Akerlof & Kranton, 2000; Lazear, 2000), we take up the broad hypothesis that incentives work best when employees have reasonable expectations that they will be successful.

Like others, we hypothesize that several factors can impact expectancy, but pay closest attention to race. In lab settings outside of the U.S., experimental analyses indicate that race/ethnicity moderates the effect of incentives on student outcomes, particularly when minoritized students' identity is made more salient (Farzana, Li, & Ren, 2015; Hoff & Pandey, 2006). We argue that race/ethnicity may play a particularly acute role in responses to incentives in the U.S. education context, where many scholars raise concerns that accountability systems lead to inequitable environments and outcomes, both for students (e.g., Au, 2016; Betts et al., 2001; Darling-Hammond, 2007) and for teachers (e.g., Campbell, 2023). Indeed, in the context of our study, Black teachers received lower evaluation ratings than White teachers, on average, leading to different consequences and outcomes.

We further examine heterogeneous effects by teaching experience (e.g., novices versus veterans), as data trends show that, for both Black and White teachers, novices performed quite differently than veterans. Disaggregating effects by experience also is consistent with theory on identify and the economics or organizations, which argues that one's standing in an organization (e.g., seniority) drives the success of incentive schemes: "outsiders" are less likely to respond to incentives than "insiders" (Akerlof & Kranton, 2000).

Finally, in supplemental analyses, we explore heterogeneity in incentive effects by the difficulty of the tasks that teachers are held accountable. Task difficulty is a longstanding way that psychologists have operationalized expectancy (Atkinson, 1957; Locke & Latham, 2002). Examining this additional dimension of heterogeneity is a useful complement for interpreting

trends related to race and experience, which we argue may also be driven by expectations of success. However, we have more limited statistical power to detect differences in effects across teaching tasks and, thus, treat these analyses as exploratory.

We apply and test our hypothesis in the context of the District of Columbia Public Schools' (DCPS) teacher evaluation system, IMPACT, which is a useful case and context for several reasons. First, it is one of the "outlier" evaluation systems with very high stakes and high fidelity of implementation (Bleiberg et al., 2023; Putnam, Ross, & Walsh, 2018). In DCPS, the lowest-performing teachers are immediately dismissed, those just above them in the performance distribution are threatened with dismissal if they do not improve the following year, and the highest-performing teachers can earn large increases in their base salary (up to $27,000 per year) if they repeat their high performance the next school year. Second, IMPACT was first implemented in 2009-10, with over ten years of district-wide data that provides sufficient statistical power to examine heterogenous responses based on teacher race and experience (and, to a lesser extent, task difficulty). Differential effects often are smaller than average effects and, thus, require larger sample sizes.

Third, public discourse has raised concerns about inequities in the system, suggesting that race may play a role in how teachers respond. IMPACT is not unique in this regard (Campbell, 2023; Steinberg & Sartain, 2021), but it is one of the highest profile examples that garnered national attention and triggered DCPS's own equity review. For example, the initial round of firings in spring 2010 were felt disproportionately by early career Black teachers, leading to protests from teachers and non-teachers at community meetings, shifts in voting blocks in city-wide elections, and lawsuits from the teachers' union. We return to and elaborate on these factors when discussing the implications of our results.

Using a regression discontinuity design that exploits strict eligibility thresholds for IMPACT's two incentives—dismissal threats and base salary increases—we, indeed, find large differences in responses at the intersection of race and teaching experience. Black novices in their first several years as a teacher were the least likely to reap the benefits of the system (i.e., most likely to be dismissed or receive a dismissal threat, and least likely to receive a salary incentive offer) and did not respond at all to either incentive. There is an obvious circularity here, as perceived "likelihood of success" and actual "success" are both defined by performance improvement. Aligned to expectancy theory, a history of differential improvement over time may inform future Black novices about their likelihood of success.

Black veterans, in contrast, were less likely than Black novices to be threatened with dismissal and more likely to be offered the salary incentive, and they responded to both incentives to a moderate degree (upwards of 0.38 teacher-level SD). While we do not estimate effects on student outcomes, literature on teacher coaching suggests that effects on similar teacher performance measures that are slightly larger in magnitude (0.5 teacher-level SD) translate into fairly large increases in student test scores (0.2 student-level SD) (Kraft et al., 2018).

White novices were less likely than Black veterans to be dismissed and similarly likely to receive a salary incentive offer; they also had the highest rates of opting into the salary incentive that required giving up some job protections. Aligned to these patterns, we find the largest dismissal threat effects on subsequent performance of this group (0.61 teacher-level SD). Further, high-performing White novices offered a salary incentive were fairly likely to receive one (41 pp increase), despite declines in performance relative to the control group who just barely missed the eligibility threshold.

While White veterans had the lowest dismissal rates and highest rates of being offered a salary incentive, they did not respond to either incentive. One explanation may be that White veterans exhibited lower expectations of success than White novices (and other groups) in more subtle ways. They were less likely than White veterans to voluntarily leave the district following a dismissal threat, potentially signaling worse (perceived) outside options. White veterans also were less likely than White novices to opt into the salary incentive when offered. Aligned to a regression discontinuity design, all our estimates are "local average treatment effects" that capture incentive effects for teachers right around the eligibility thresholds, rather than capturing the net effects of the IMPACT program as a whole.

These patterns paint a more nuanced story of IMPACT and its effects compared to several prior analyses. Focusing on the first couple of years of implementation and using a similar regression discontinuity design, Dee and Wyckoff (2015) found improved performance both for low-performing teachers threatened with dismissal (0.27 SD) and for high-performing teachers offered the financial incentive (0.24 SD). A follow-up study identified sustained effects through six years of implementation (Dee et al. 2021), though this analysis only focused on dismissal threats and not the salary incentive. Our findings suggest that these average effects across teachers mask large differences by teacher race and experience. To our knowledge, this is the first study to causally examine heterogeneous responses to teacher evaluation incentives along these dimensions, though DCPS and IMPACT certainly are not the only contexts where equity concerns are raised (Campbell, 2023; Steinberg & Sartain, 2021).

In the conclusion, we discuss the need to create a more level playing field where teachers from different backgrounds feel similar expectations of success, which may help teacher incentive schemes function more as intended.

Conceptual Frameworks and Motivating Literature

Education policy often has explored accountability systems as a mechanism for improved individual- and system-wide performance, but in doing so has raised substantial concerns related to equity. Standards-based reforms, in which student test scores are a key metric for school performance, have led to several perverse incentives and concerns for racially/ethnically minoritized students. Applying Omi and Winant's (2014) concept of "racial projects" to the context of standards-based reform, Au (2016) argues that high-stakes standardized testing is often publicized as a tool for achieving racial equity. Because of their purported objectivity, high-stakes tests can help school systems identify under-resourced schools and communities, and then allocate resources to them. Yet, in practice, Au argues, these accountability systems have negatively impacted students of color by narrowing the curriculum and pedagogy, and disciplining Black and Brown students. Other scholars raise similar concerns (e.g., Darling-Hammond, 2007; Betts et al., 2001), further pointing out that high-stakes testing can lead schools to game the system by excluding lower-performing students (often Black and Brown) from testing and from school (Heilig & Darling-Hammond, 2008).

Though newer and smaller, the literature on accountability for teachers raises similar equity concerns. For example, when teachers are evaluated based on student performance and on standards-based rubrics, they may be pushed to align (and narrow) the curriculum, leaving little room for teachers of color (and others) to engage in culturally responsive teaching (Achinstein & Ogawa, 2012). Another key concern for scholars and for teachers is that subjective measures of teacher performance, including classroom observations, may be racially biased. Using statewide data from North Carolina's teacher evaluation system, Campbell (2023) shows that Black women received lower classroom observation ratings than White women, even when controlling for a

second measure of teacher effectiveness: their contributions to student test-score growth. In turn, Black women were two times as likely to be placed on a punitive professional development plan than White women. Steinberg and Sartain (2021) find similar differences between Black and White teachers in Chicago's teacher evaluation system, though show that much of the gap in observation scores is driven by sorting of teachers into schools. In a research study where school leaders rated teachers' instruction but without high stakes attached, Campbell and Ronfeldt (2018) find similar race-based differences across six additional districts. Potentially driven by these patterns, a nationally representative sample of teachers described teacher evaluation systems as fair for themselves (close to 90% agreed) but less fair for all teachers in their school (roughly two-thirds agreed) (Tuma, Hamilton, & Tsai, 2018).

While education research provides substantial evidence and discussion on the fact that teacher evaluation systems—like education accountability systems more broadly—are racialized, there is much less evidence on how individuals respond as a result. Theoretical and empirical work from other research areas and disciplines provides helpful insight. Economics is an appropriate starting point, as education policies focused on standards-based reform, accountability, and evaluation largely stems from economic theory. More specifically, subfields in contract and personnel economics are built on the fundamental premise that performance incentives motivate employees to change their behavior to maximize personal utility, ultimately increasing firm output (Lazear 2000; Holmström, 1979). Merit pay for teachers—which necessarily include an evaluative component—is widely explored in both theoretical and empirical investigations on this topic (Borjas, 2020; Ehrenberg & Smith, 2016; Holmström & Milgrom, 1991; Lazear, 2003). A growing literature base supports this theory when performance incentives are implemented across job

sectors (Weibel, Rost, & Osterloh, 2010). At the same time, economists also point out how the theory can break down (Jacob &Levitt, 2003) and raise questions about its economic value.

For example, Hoff and Pandey (2006) and Farzana, Li, and Ren (2015) find 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. In India and rural China, the authors randomly assigned students to tasks that, if completed successfully, had a financial reward. When minoritized students were primed about their identity, they performed worse. Within economics, these findings uphold Akerlof and Kranton's (2000, 2005) social identity theory, which broadly argues that "outsiders" need larger monetary incentives than "insiders" to compensate them for acting in the interest of the firm rather than their own. When offered the same incentive, outsiders are less likely to respond than insiders. Outsider versus insider status can be influenced by hierarchies specific to an organization (e.g., seniority in the firm), or to hierarchies within larger global and social networks (Lloyd & Mertens, 2018).

Hoff and Pandey's (2006) and Farzana, Li, and Ren's (2015) findings also have clear connections to psychology. Both sets of authors hypothesize that one likely explanation for their results relate to 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, & Aronson, 2002). A related explanation for these patterns is that experiences of marginalization among certain groups affect individual goals, decrease expectations of success, and provide less motivational value (Eccles & Wigfield, 2020). Together, theoretical and empirical literature from multiple disciplines highlight the fact that responses to incentives are very likely to be heterogeneous, driven by race/ethnicity, seniority, and other factors that shape one's sense of self and expectations of success.

Circling back to the context of teacher evaluation systems, there is some existing evidence that responses to incentives may be moderated by expectations of success. Also focusing on DCPS, Phipps and Wiseman (2021) find that teachers tend to improve in their performance most when they anticipated an upcoming classroom observation, and that these improvements were driven by "easier" tasks related to classroom procedures and routines compared to "harder" tasks related to instructional rigor. However, to our knowledge, no studies have examined effects of teacher evaluation incentives by race and seniority that likely shape expectations of success.

Performance Incentives for Teachers in DCPS

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 base salary increases.[1] Both incentives align to assumptions of rational, utility-maximizing employees who aim to keep their job and earn the highest possible salary (Lazear, 2000; Holmström, 1979). The evaluation and incentive system still operates today with the same basic structure we describe below (see online Appendix Table 1 for an overview of changes in the evaluation system design across years). However, we focus our analyses on school years through 2018-19, given Covid-related interruptions in performance monitoring and incentive rollout starting in the 2019-20 school year.

Our paper focuses on heterogeneous responses primarily by race and teaching experience. For race, we disaggregate effects for Black teachers (50% of the DCPS workforce) and White teachers (32%; see Table 1). Teachers from other racial/ethnic backgrounds each made up less

[1] Like other teacher evaluation systems (Bleiberg et al., 2023), IMPACT includes mechanisms other than "carrot and stick" incentives as means of shifting system-wide teacher quality and student outcomes. The evaluation system provides an avenue for clarifying a vision of excellent instruction and providing feedback to teachers about how to improve their practice (Phipps & Wiseman, 2021). Though not a focus of this paper, the district's theory of change also posits that the redistribution of teachers, including replacing low-performing teachers with higher-performing ones, benefits students (Adnot et al., 2017).

than 5% of the DCPS workforce. For teacher experience, we further disaggregate effects for novices in their first four years in the job versus veterans, each of whom comprised roughly 50% of the teacher workforce in DCPS (see Table 1). In addition to maximizing sample sizes for both groups, this division aligns with literature showing that performance improvement trajectories are quite steep in the first four years and then taper off after that (Papay & Kraft, 2015; Rockoff, 2004).

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 school year and the availability of other metrics); (ii) student achievement growth on a district-administered assessment (up to 50% for teachers who work 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%); (v) school-aggregated student test-score performance (up to 5%, but only in the first three years); and (vi) students' assessment of teachers' practice (up to 10%, but only starting in the seventh year).

Measures (ii) through (vi) were captured once yearly. Depending on their prior-year evaluation score, teachers were observed up to five times each year: up to three times by an administrator from teacher's own school (often the principal), and up to two times by a content area expert employed by the district expressly for the purpose of conducting evaluations (called a "master educator"). When the district changed observation rubrics in the 2016-17 school year—

from the Teaching and Learning Framework (TLF) to Essential Practices (EP)—it also removed the requirement that teachers be observed by a master educator.[2]

The summary measure of teacher performance—which is a weighted average of the individual components—ranges from 100 to 400, with multiple thresholds that determine allocation of incentives (see Figure 1). At inception in the 2009-10 school year, 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"—in part to address rising district-wide performance. Over a decade, the median IMPACT score shifted from roughly 300 to 340, roughly equivalent to 1 SD. Further, the performance distribution developed bimodal peaks in the "Effective" and "Highly Effective" bands. Importantly, though, distributions are smooth across the thresholds that trigger incentives (see below for formal density tests).

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 & Wyckoff, 2015; Whitmire, 2011). In practice, though, only 15% of teachers worked in a grade and subject where district-wide testing is 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 67% of the overall score. For over 90% of teachers, observations accounted for over 50% of their performance score. On average

[2] On average, teachers in our sample were observed four times per year. In any given year, most teachers (71%) were observed multiple times by the same school leader. When the district also relied on master educators, roughly half of teachers (53%) were observed by just one outside rater and the rest were observed by two different master educators. Within each school and year, an average of three different school leaders and 15 master educators conducted observations.

across teachers, the summative IMPACT score is correlated with the observation component at 0.9, highlighting the weight that observations play. Because observation scores comprised the majority of the total evaluation score for the majority of teachers, it is intuitive that teachers' behavioral responses focus on tasks and skills identified in the rubric. In our main analyses, we focus on a summary teacher observation score, averaged across lessons, as our key measure of teacher performance (adjusted teacher-year intra-class correlation [ICC] = 0.79). In exploratory analyses, we disaggregate observation scores by sub tasks related to the classroom environment, organization of the lesson, and rigor of the content.

Prior research and DCPS's own equity review suggest that classroom observations can be subjective and raise concerns about potential racial biases amongst raters. We do not have a definitive way to identify or rule out racial bias, which would require having an objective or "true" measure of teacher performance and then seeing how raters' scores differ depending on teachers' race/ethnicity. However, like the prior work (Campbell, 2023; Campbell & Ronfeldt, 2018; Steinberg & Sartain, 2021), we provide several measurement tests and checks. In Table 2, we show differences in teachers' lesson-level classroom observation scores by race/ethnicity and explore possible sources of the gaps. The raw difference between Black and White teachers is quite large (0.3 SD; column 1). Accounting for teaching experience (column 2) exacerbates the gap to some extent (0.34 SD). Conversely, including school fixed effects to account for sorting of teachers and differences in raters across schools shrinks the gap in half (0.18 SD) but does not close it.

Next, we examine whether teacher-rater race/ethnicity matches are correlated with higher (or lower) scores (column 4). Black teachers score 0.06 SD higher when they have a race-matched rater, while there is no difference for White teachers. Patterns are very similar when we replace school fixed effects with teacher fixed effects (column 5), making within-teacher comparisons

across lessons scores by a race-matched versus non-race matched rater. Race-matching effects, for Black teachers, is a likely signal of some degree of racial bias: two different raters, of different races, scored the same teachers' instructional quality differently. Notably, though, 60% of Black teachers' lessons were scored by a Black rater, while roughly one-third of White teachers' lessons were scored by a White rater. Thus, race-matching and associated biases likely explains only a small portion of the overall Black-White difference in teacher observation scores.

We extend these analyses in online Appendix Table 2 by examining the extent of within- and across-school sorting of raters to teachers. Non-random sorting could be another driver of race-based differences in observation scores. To do so, we present ICCs from random-effects models that estimate the proportion of variance in baseline teacher characteristics (e.g., prior-year observation score, race/ethnicity, experience) that lies within versus across raters. Overall, we find minimal evidence of sorting of teachers to raters based on observable teacher characteristics. Through the 2015-16 school year, teachers were observed by two types of raters: school leaders and master educators hired by the district specifically for the purpose of evaluation. Because school leaders only observed teachers from their schools, we estimate ICCs conditioning on school fixed effects. For master educator sorting tests, we estimate ICCs including and excluding school fixed effects. We also estimate ICCs that condition on school-year, leave-out averages of the baseline teacher characteristic, in order to account for the fact that schools differ in their staff characteristics. In most instances, ICCs are below 0.02 and often zero (to three decimal places).

Together, the evidence provides possible but not definitive nor extensive evidence of rater racial biases. Like other contexts, much of the differences in observation scores between Black and White teachers are related to schools.

The Incentive Structure: Mapping Rollout and Take-Up to Expectations of Success

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 because of their "Minimally Effective" rating 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). Starting in the 2012-13 school year and aligned to the creation of a new performance band ("Developing"), two additional sets of teachers were threatened with dismissal based on different combination of low scores: one "Developing" rating followed by one "Minimally Effective" rating, or three consecutive ratings below "Effective". Below we describe sample restrictions to ensure a clean treatment-control contrast across all school years.

Roughly 2% of all teachers received an "Ineffective" rating meant to lead to immediate dismissal and 4% of teachers received a "Minimally Effective" rating that resulted in a dismissal threat. Of those teachers who were threatened with dismissal, 36% voluntarily left the district. Of the teachers who remained, 23% were dismissed the next school year. That DCPS actually dismissed low-performing teachers stands in sharp contrast to most other school districts, where almost everyone receives a satisfactory rating (Kraft & Gilmour, 2017; Weisberg et al., 2009).

Rates of separation and dismissal varied substantially based on teachers' race and experience. Black novices were the most likely to experience the consequences of the evaluation system (3% immediately separated and 6% threatened with dismissal), and White veterans were

the least likely (less than 1% immediately separated and 1% threatened with dismissal; see Figure 2). Black veterans and White novices had similar likelihoods of immediate separation (1%) or being threatened with dismissal (3%). At the same time, White novices threatened with dismissal voluntarily left the district at higher rates than other groups (40%, compared to 28% for White veterans, 29% for Black novices, and 31% for Black veterans; see online Appendix Table 3), potentially signaling better (perceived) outside options. In turn, White novices threatened with dismissal who stayed in the district were the least likely to be separated afterward (13%, compared to 23% for Black veterans and 24% for both Black novices and White veterans). Although these patterns are purely descriptive, they suggest that—among low-performing teachers—Black novices potentially had the lowest expectations of success and White novices and veterans the highest.

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, as well as a permanent increase in base pay if they received this same rating the following school year. Opting into this portion of the evaluation system required teachers to 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. Almost two-thirds of eligible teachers scoring "Highly Effective" opted into IMPACT*plus* (see Figure 2).

Base-pay increases 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. Education and experience determine base salary in DCPS and in most other school districts across the U.S. (Hanushek, 2007).

Base salary increases also depended on the poverty level of teachers' schools. Teachers in high-poverty schools—defined by the district as schools where 60% or more of students were eligible for free or reduced-price lunch—could receive the maximum salary incentive. In the first several years of implementation, teachers in low-poverty schools were eligible for a slightly smaller base-salary increase: a boost on the salary schedule of three rather than five experience levels, but the same jump for degree level. Starting in the 2012-13 school year, teachers working in low-poverty schools no longer were eligible for any base salary increase, though they were eligible for one-time bonuses. Also starting in the 2012-13 school year, the district implemented a career ladder that was layered on top of the performance ratings, where both "Effective" and "Highly Effective" teachers were eligible for base salary increases as they advanced from the base ("Teacher") to the top rung ("Expert Teacher"). Above the middle rung of the ladder ("Advanced Teacher"), only two consecutive "Highly Effective" ratings triggered the salary incentive.

Following similar patterns for race- and experience-based gaps in dismissal threats, Black novices were the least likely to receive a salary offer (6%) and White veterans were the most likely (15%), though not everyone accepted it. We interpret these differences in incentive rollout and take-up as signaling different expectations of success, arguing again that Black novices likely had the lowest expectations of success. That said, mapping social identity markers to expectations of success is not strict, and there are complications to the trends we describe above. For example, contingent on receiving a salary offer, Black and White novices opted in at similar rates to each other (76%), and at substantially higher rates than Black and White veterans (51% and 58%; see Figure 2 and online Appendix Table 3). Because opting in required teachers to give up some job protections, these patterns could be interpreted as a signal of expectations of success and, relatedly, risk tolerance (Prendergast, 1999). Veterans already embedded in the school system—with some

close to retirement—may not be willing to give up job security. As such, we argue that White novices may have had the highest expectations of success, even though White veterans generally performed better. Opting in may also be a signal of the value teachers place on monetary incentives (Gneezy & Rustichini, 2000). Nonetheless, the descriptive patterns overlap to a large degree with expectancy theory (Atkinson, 1957) and identity economics (Akerlof & Kranton, 2000), and lead us to test for heterogeneous effects of the incentives by race and experience.

Empirical Strategy, Data, and Sample

To estimate the causal effect of dismissal threats and salary incentives on subsequent teacher behavior and 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 & Wyckoff, 2015; Dee, James, & 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.[3]

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

$$Y^{m}_{is(t+1)} = \alpha + \beta_1 * 1(S_{it} \leq 0) + f(S_{it}) + \beta_2 * 1(S_{it} \leq 0) * f(S_{it}) +$$

[3] 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.

$$\pi_t + \gamma X_{it} + \delta Z_s + \varepsilon_{it} \qquad (1)$$

where Y is a measure of incentive consequences, rewards, or performance 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 full set of outcomes, m, include: voluntarily leaving the district the following year (i.e., left but not formally separated), separation from the district following a threat (relevant to the dismissal threat sample), receiving a salary increase following the offer (relevant to salary incentive sample), and performance on the classroom observation rubric.

The parameter β_1 reports the effect of receiving an initial "Minimally Effective" or "Highly Effective" rating on a given outcome. By fitting a polynomial of the forcing variable that determines eligibility for each incentive on either side of the threshold, we can estimate the "jump" at that threshold. For dismissal threats, the treatment group falls below the threshold, i.e., $S_{it} \leq 0$ in equation (1). The reverse is true for the salary incentive, i.e., $S_{it} \geq 0$, and we adapat equation (1) accordingly. Empirical tests support a linear function of the forcing variable when estimating effects of dismissal threats and a quadratic function for effects of salary incentive. Because the choice of polynomial and the window within which the "jump" is estimated are critical for identification (Cattaneo & Titunik, 2022), in a set of robustness tests we re-estimate effects varying the functional form with high-order polynomials, reducing the bandwidth on either side of the threshold (from our preferred window of 50 to 40, 30, and 20), and implementing local polynomial estimators with robust bias-corrected confidence intervals. To increase precision, we control for baseline teacher characteristics, X_{it}, and school characteristics, Z_s (see Table 1). Because we pool data across all school years, we include year fixed effects, π_t. Finally, ε_{it} is a mean-zero error

term, and robust standard errors are reported to account for heteroskedasticity. We cluster standard errors at the teacher level, as teachers can show up in the sample across multiple school years.

To estimate heterogeneous responses to incentives by subgroups of teachers, we interact treatment indicators and the forcing variable function by dummy indicators for each of four subgroups: Black novices, Black veterans, White novices, and White veterans. We include dummy variables for the relevant subgroups in the vector, X_{it}, so that the control-group mean is estimated separately for each group. We exclude the small subset of teachers who are neither Black nor White, or missing data on race/ethnicity or teaching experience (see Table 1). Estimating one model with interacted treatment effects by subgroup allows us to directly and easily test between-group differences in coefficients. We show below that patterns of results are the same if we run models separately for each subgroup in unstacked data.

All estimates should be interpreted as intent-to-treat, local average treatment effects that are generalizable to individuals close to the incentive threshold. For dismissal threats, the intent-to-treat is the same as treatment-on-the-treated because no teachers could opt out of this incentive. In contrast, teachers could opt out of the salary incentive, and so readers interested in treatment-on-the-treated can scale-up the intent-to-treat estimates by the take-up rate (see Figure 2). We do not use two-stage least-squares techniques—with the salary incentive offer as an instrument for the opt-in decision—as we lose statistical power to be able to detect heterogeneous responses to this incentive, which is the main purpose of the analysis.

To identify our analytic samples, we start with the full population of DCPS teachers: roughly 3,500 individuals per year, between 2009-10 and 2017-18.[4] We also rely on data from the 2018-19 school year, but only for capturing outcomes triggered by incentive eligibility in the prior

[4] We exclude a small fraction of teachers devoted to supporting students with special education needs, as these schools often used a distinct classroom observation rubric from other teachers.

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. (Focusing only on the first several years of implementation, Dee and Wyckoff [2015] used bandwidths up to 100 points.) 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.

Next, we make several restrictions to ensure clean treatment-control contrasts. For the dismissal threat sample, we drop teachers slated for separation at the end of the current year because of their past performance (e.g., teachers in their second year with a "Minimally Effective"). These teachers automatically are dropped from analyses that look at effects on next-year performance outcomes, as these scores are not available. Removing them from all analyses allows us to examine a consistent base sample, as well as to avoid conflation of voluntary and involuntary leaves. In data from 2012-13 onward, we also exclude the set of potential control-group teachers who received a "Developing" rating in the current year and a "Developing" or "Minimally Effective" rating in the prior year, as this group also was up for dismissal the following year if they did not improve. Therefore, there is no treatment-control contrast. Following Dee and colleagues (2021), we exclude the 2009-10 school year from dismissal threat analyses given anecdotal discussion with DCPS leadership and empirical evidence that the dismissal incentive was not yet fully implemented (Dee & 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 other years. Like Dee, James, and Wyckoff (2021), we interpret estimates from the remaining sample and resulting treatment-control contrast as the "credible and immediate" (p. 315) dismissal threat for receiving a "Minimally Effective" rating. In the first several years of implementation, the control group received no threat, while in later years the control group faced a dismissal threat but had more time to improve.

For the salary incentive analysis sample, we exclude teachers in low-poverty schools after the 2012-13 school year because this group no longer was eligible for a base salary increase. Following the creation of the career ladder in the same year, we also exclude teachers below the middle rung, where both "Effective" and "Highly Effective" teachers were eligible for a base-pay increase. Finally, we exclude teachers who already received a salary increase in a prior year. While teachers are able to receive multiple salary increases, behavioral responses likely differ after receiving one. Because all teachers were incentivized to improve over time to earn consecutive top performance ratings, we interpret the treatment-control contrast of "Highly Effective" versus "Effective" teachers similarly to the dismissal threat sample: we estimate the credible and immediate offer of a base salary increase for maintaining the top performance rating. These sample restrictions echo the technique of frontier RD, which uses multiple variables to determine assignment to treatment (Reardon & Robinson, 2012; Wong, Steiner, & Cook, 2013).

RD designs have a strong causal warrant (Campbell, 1969; Lee & Lemieux, 2009), but like any design have embedded assumptions. An important concern with any RD design is that there may be systematic sorting across the performance threshold. If teachers—or raters—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 & Lemieux, 2008; Lee & Lemieux, 2009; McCrary, 2008), and empirical examination in our data suggests that the assumption holds. In Table 3, we look for jumps in background teacher and school characteristics at the incentive thresholds by specifying versions of equation (1) that replace teacher outcomes with these baseline characteristics. While some individual estimates are statistically significant, we cannot reject the null on a joint test of significance ($p = 0.447$ for the dismissal threat sample and 0.208 for the salary incentive sample). We also fail to reject the null hypothesis of a smooth distribution of the running variable across the eligibility threshold ($p = 0.694$ and 0.331), using the local polynomial density estimator proposed by Cattaneo, Jansson, and Ma (2020).

It is important to note that several of our outcome measures are available only for teachers who returned to the district the next year. (The exception is the voluntary leave measure, which we can measure for all teachers.) While differential attrition from the sample could lead to imbalanced groups, our baseline balance and density tests hold in the subsample of teachers who are observed in the data in year $t + 1$ (see Table 3). Earlier, we presented evidence that some groups (i.e., White novices) were more likely than others to leave the district following a dismissal threat. This descriptive analysis focused on the treatment group only. The fact that balance remains intact in year $t + 1$ indicates that the characteristics of teachers are smooth across the eligibility threshold, comparing teachers who were/were not threatened with dismissal and those who received/did not receive a salary incentive.

That said, it may still be that the teachers who stay are more optimistic (or have private information) about their likelihood of success in year $t + 1$, which we cannot observe. They may also be in contexts where they have reason to be more optimistic. For example, teachers below the threshold in year $t + 1$ are significantly less likely to be in high-poverty schools. This is all

consistent with the storyline of the paper related to expectancy. However, it could affect the interpretation of results, where the effect is not necessarily coming entirely from an increase in effort but partly due to selection out. At the same time, we show below that some groups—like Black novices—do not see an increase in subsequent performance, which provides less support for the selection out mechanism.

Results

To begin, in Figure 3, we show graphical evidence of the impact of dismissal threats (Panel A) and salary incentives (Panel B) on subsequent outcomes, on average across teachers. The first two figures in each row show impacts on binary measures capturing incentive consequences and rewards: voluntary leave, separation (for dismissal threat sample only), and base-pay increase (for salary incentive sample only). The third figure in each row shows effects on the summary classroom observation score, in teacher-level SD units. In all figures, the *x*-axis is the current-year IMPACT score, which serves as the forcing variable that determines eligibility of incentives. Here, we keep the IMPACT score in its original scale to examine discontinuities at the relevant thresholds (i.e., 250 for dismissal threats, 350 for salary incentives).

Visually, the graphs show jumps in incentive consequences and rewards at the eligibility threshold: roughly 7 percentage point increase in the likelihood of voluntarily leaving the district and 4 percentage point increase in the likelihood of being separated from the district for the low-performing teachers following a dismissal threat, relative to slightly higher-performing teachers at baseline who just missed this incentive (i.e., jump to the left of the threshold); and roughly 37 percentage point increase in the likelihood of earning a salary increase for high-performing teachers following the offer, relative to slightly lower-performing teachers who barely missed this incentive (i.e., jump to the right of the threshold). For dismissal threats, we also observe a jump in

the classroom observation score (0.22 SD), indicating that classroom teaching performance increased, on average, across teachers. One point of comparison for interpreting this standardized effect size is the difference in average performance between novice and veteran teachers (0.35 SD). We do not see any meaningful difference in performance for teachers offered a salary increase. One immediate conclusion is that salary incentives are a less potent incentive than dismissal threats, or that it may be harder to improve from a high baseline score versus a lower one.

On their own, these findings suggest that the dismissal threats worked to improve subsequent teacher performance, while the salary incentives did not (even though teachers still were fairly likely to receive a salary increase). These findings also are consistent with prior work that uses the same research design, but with fewer years of data (Dee & Wyckoff, 2015; Dee et al., 2021). At the same time, these average effects mask some differences in effects by teacher race and experience.

Heterogeneous Incentive Impacts by Race and Experience

Next, in Table 4, we report regression estimates of the effect of dismissal threats and salary incentives, on average across teachers (which align with findings from Figure 3) and by subgroups of teachers: Black novices, Black veterans, White novices, and White veterans. We show visual confirmation of discontinuities in outcomes by the four subgroups in online Appendix Figure 1. At the bottom of each table, we further report *p*-values on tests of coefficient equivalence between the four subgroups of teachers based on race and experience. For visual ease of interpretation for readers, we bold *p*-values below 0.1. We set a slightly higher threshold for statistical significance given that between-task or between-group differences generally require higher statistical power compared to tests of null hypotheses that individual coefficients are different from zero. At the

same time, statistical tests on multiple outcomes and multiple subgroups could lead us to observe a false positive due to chance alone. Therefore, we also consider a Benjamini-Hochberg (1995) adjustment that accounts for the number of tests conducted (n = 116) and an allowable false discovery rate, which we set at 20%. The resulting critical value for statistical significance is 0.058.[5]

For dismissal threats, we observe the largest effects on the subsequent classroom teaching performance of White novices (0.61 SD). We also estimate positive, statistically significant, and economically meaningful effects of dismissal threats on the performance of Black veterans (upwards of 0.38 SD). The latter effect is roughly two-thirds the former, but not statistically significantly different. For Black novices, we find no effect of dismissal threats on subsequent performance, with a point estimate right around zero. For White veterans, we observe a negative point estimate that is not statistically distinguishable from zero given the small share of White veterans threatened with dismissal. That said, these null—and potentially negative—effects for Black novices and White veterans are consistently distinguishable from the positive effects for Black veterans and White novices.

Between-group differences in dismissal threat effects on performance naturally translate into some differences in the likelihood of being separated. For example, dismissal threats lead to lower effects on separation for White novices (-4 percentage points [pp]) compared to White veterans (15 pp; p = 0.064 on difference). Similarly, for Black novices, lack of improvement in performance may also lead to greater likelihood of separation (6 pp), though the effect is insignificant. Point estimates for voluntary leave rates vary to some extent across subgroups,

[5] Following Benjamini-Hochberg (1995), we start by rank ordering all p-values from smallest to largest. To calculate the critical value for each p-value, we use the formula $i/m*Q$, where i is the rank of the p-value, m is the total number of tests (116 in our case), and Q is the false discovery rate that we set at 0.2. Finally, we identify the largest p-value that is less than the adjusted critical value.

though none of the between-group differences are statistically significant. Descriptive analyses show that White novices threatened with dismissal had very high voluntary leave rates (see Figure 2). However, the control group of White novices who just barely missed receiving a dismissal threat also left at similarly high rates. In contrast, Black veterans threatened with dismissal had lower voluntary leave rates than White novices (see Figure 2), while the corresponding control group had even lower voluntary leave rates (12 pp). The dismissal threat effect on voluntary leave rates for Black veterans is statistically significantly different from zero.

For salary incentives, quick examination of *p*-values reveals few between-group differences, as well as null effects on performance outcomes for most groups (which is similar to visual findings from Figure 3). However, there are several notable patterns. We find evidence that White novices declined in their classroom teaching performance relative to the control group (-0.18 SD). Further, all four race-by-experience subgroups were fairly likely to earn a salary increase the following year, from 43 pp increase for Black veterans to 44 pp for Black novices. If we adjust the intent-to-treat salary receipt effects by the take-up/opt-in rate (see Figure 2), then salary receipt rates are even higher: 58 pp for Black novices, 66 pp for Black veterans, 54 pp for White novices, and 67 pp for White veterans. That all groups still received salary increases despite no increases in performance—and possible declines for White teachers—reflects the incentive system's requirement for *maintaining* their "Highly Effective" rating the following year, rather than *improving* in their overall score.

It also is possible that high baseline scores for high-performing teachers offered a salary increase may create ceiling effects. We explore this possibility below by disaggregating effects by teaching tasks, which vary in the degree of room for improvement.

Exploratory Analyses Related to Teaching Task

Another possible dimension of heterogeneity in response to evaluation incentives is teaching task—and more specifically, task difficulty—which is the primary way that psychologists have operationalized expectancy dating back to at least the 1950s (Atkinson, 1957). Because we interpret differential responses to evaluation incentives by race and experience through an expectancy lens, we view supplemental analyses related to task as a useful complement. Is there evidence that expectancy drives incentive-based responses across the multiple dimensions of heterogeneity we explore?

Teaching and teacher evaluation systems are an interesting case to consider differential responses by task for several reasons. Teaching is a multidimensional, multi-task job that requires teachers to make decisions about where to allocate their time and attention. In fact, teaching is a key illustrative example in discussion of the multitask "principal-agent" problem in economics, which is concerned with contract setting for multitask and multidimensional jobs. Holmström and Milgrom (1991) describe how teachers are expected to improve student achievement on standards-based assessments, as well as to promote curiosity and creative thinking, build interpersonal relationships with students, and manage 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. Further, while a concern in the principal-agent problem is that teacher performance cannot always be monitored well, modern day teacher evaluation systems do in fact include multiple performance measures that vary substantially in their focus and difficulty.

To begin this analysis, we start by decomposing the summary classroom observation score into sub components and tasks. In the first several years of implementation, DCPS used the Teaching and Learning Framework (TLF) observation instrument that included nine items. In 2016-17, the district switched to a new tool called Essential Practices (EP) with five items. For the sake of parsimony and to accommodate changes in the observation rubric over time, we create three composite measures of task performance.[6] Aligned to Atkinson's (1957) definition of objective task difficulty, in online Appendix Table 4, we arrange these tasks based on average performance across all DCPS teachers from highest to lowest scores, or easiest to hardest: (i) *Build a Supportive Classroom* assesses the extent to which teachers engage all students in the learning activities and invite students back in when they become disengaged (mean = 3.43 out of 4; adjusted teacher-year ICC = 0.65); (ii) *Lead Well-Planned and Responsive Lessons* assesses the extent to which teachers maximize instructional time, create and lead linked learning activities, explain content clearly, and check and respond to student (mis)understanding (mean = 3.22; ICC = 0.75); and (iii) *Engage Students in Rigorous and Higher-Level Work* assesses the extent to which content is aligned to grade-level standards, is intellectually challenging, and maximizes students'

[6] Before the launch of the evaluation system in the 2009-10 school year, a team of teachers, school leaders, and central office staff created a rubric to score the quality of teachers' classroom instruction called the Teaching and Learning Framework (TLF), which drew from instructional research from several other observation instruments (Danielson, 2011; Pianta & Hamre, 2009; Wiggins & McTighe, 2005). A key goal was to create a common language to discuss teaching and learning and for providing clear expectations for teacher performance (DCPS 2010). This original rubric contained nine total tasks and items, some with subcomponents. At the end of the first year of implementation, the district streamlined the rubric to nine tasks without subcomponents. In the 2016-17 school year, the district commissioned the development of a new rubric called Essential Practices (EP), with five dimensions that overlap to a large degree with the original nine. In online Appendix Table 4, we list and provide definitions for all dimensions from each rubric, as well as create a crosswalk between rubrics. Notably, rank ordering by average teacher performance of similar tasks and dimensions is the same. Dimensions focused on building a supportive classroom receive the highest scores and tasks focused on higher-level understanding receive the lowest scores. Average scores on the EP rubric are higher than on the TLF, driven by district-wide increases in observation scores and performance over time (see Figure 1). We further reduce dimensionality by creating a parsimonious set of three tasks. *Build a Supportive Classroom* includes one item that is defined very similarly in both the TLF and EP rubrics. *Lead a Well-Planned and Responsive Lessons* includes six items from TLF (internal consistency reliability [α] = 0.87) and three items from EP (α = 0.8). *Engage Students in Rigorous and Higher-Level Work* includes two items from the TLF (α = 0.74) and two items from EP (α = 0.7) that are worded very similarly. The three measures are highly correlated (r = 0.66 to 0.83) though represent fairly distinct components of teachers' work.

ownership of learning (mean = 2.95; ICC = 0.66). Rank ordering is consistent with other studies where metrics relating to questioning technique often are deemed the hardest and metrics related to the classroom environment the easiest (Hamre et al., 2013; Kane & Staiger, 2012).

We examine differential responses to evaluation incentives by task in Figure 4, where we include estimates across all DCPS teachers and for the four race-experience subgroups. Regression estimates and formal tests of coefficient equivalence are presented in online Appendix Table 5. On average across teachers, we find some evidence of differential effects of dismissal threats by teaching task. The estimated dismissal threat effect on the highest scoring task, potentially with the highest expectation of success (0.25 SD for *Build a Supportive Classroom*) is almost twice as large as the effect on the lowest-scoring task, potentially with the lowest expectations of success (0.14 SD for *Rigorous and Higher-Level Work*; $p = 0.054$ on difference between these effects). For race-experience subgroups, magnitudes of effects on each of the three teaching tasks generally go in the expected direction, potentially justifying our interpretation of differences in incentive response across race-experience groups as related to expectancy. However, differences in effects between tasks are not statistically significant when disaggregated by subgroup.

For salary incentives, the evidence regarding heterogeneity in effects by task is less clear, largely because most teachers did not improve in their average performance in response to this incentive. For example, we previously documented declines in average performance amongst White veterans offered a salary increase, relative to the control group, and these declines are fairly similar across the three teaching tasks. However, for Black veterans, disaggregating effects by task reveals new insight. For this group, we find positive and statistically significant effects of salary incentives on *Lead Well-Planned and Responsive Lessons* (0.17 SD). Further, Black veterans improved more on this task relative to the two other tasks—one harder (*Rigorous Work*; $p = 0.001$)

and one easier (*Supportive Classroom*; $p = 0.057$). This pattern may initially seem counter to expectancy theory, which posits larger incentive effects on easier tasks relative to harder ones. At the same time, a likely reason is that very high-performing teachers offered a salary increase already were excelling in the easiest task with little room for improvement (mean = 3.8 on a scale from 1 to 4). In other words, high-performing Black veterans offered a salary incentive improved their performance on the easiest task where there still was room for growth.

Figure 4 helps illustrate that, while there are some differences in responses to incentives by task, the differences between race-experience groups tend to be much larger. With more years of IMPACT data, we may be able to detect statistically significant differences in effects by task (as well as to increase precision in dismissal threat effects for White veterans). Power calculations using the *PowerUp!* tool (Dong & Maynard, 2013) indicate that, in the current sample, we have the power to detect effects of dismissal threats as small as 0.16 SD using the full sample (and 0.3 to 0.7 SD for race-experience subgroups); for salary incentives, we have the power to detect effects as small as 0.11 SD using the full sample (and 0.17 to 0.32 SD for race-experience subgroups).

Robustness Tests

We conduct a variety of robustness tests to ensure the internal validity of our results. Here, we focus on the outcome measures explored in our main analyses. We also exclude formal tests of coefficient equivalence between subgroups and teaching tasks, as the robustness tests generally decrease power and precision.

First, in online Appendix Tables 6, we show that results are qualitatively similar when we replace observable school characteristics with school fixed effects, thus comparing teachers across schools. In other words, sorting of teachers to schools, differences in rater pools within schools, and other school-specific contexts do not appear to drive our results. In the bottom panel of the

same table, we keep observable school characteristics and vary the functional form of the forcing variable to ensure that the "jump" in outcomes at the threshold is of the correct sign and magnitude. Whereas our primary results include a linear function for dismissal threats and a quadratic function for salary incentive, here we increase the order of the polynomial by one (i.e., quadratic and cubic). Patterns of results lead to the same conclusions.

In online Appendix Table 7, we report estimates that restrict the bandwidth to an increasingly narrow range around the performance thresholds, which decreases the model's reliance on functional form assumptions. For bandwidths of 40, 30, and 20 points on either side of the threshold, patterns of results remain relatively stable, confirming that observations far from the performance threshold are not driving our conclusions. At a bandwidth of 20, we observe a large dismissal threat effect on the voluntary leave rates of White veterans. However, the sample size is quite small, given how few White veterans were threatened with dismissal. The direction of estimates are the same at larger bandwidths.

Finally, in online Appendix Table 8, we ease the parametric form assumption by re-estimating results with a local polynomial regression discontinuity estimator that offers robustness to large bandwidth and produces bias-corrected confidence intervals and inference procedures (Calonico, Cattaneo, & Titiunik, 2014). Expectedly, standard errors are slightly larger, but point estimates and overall patterns of results are quite similar to the main results presented earlier. The local polynomial estimator requires estimation of effects in subgroup samples. Therefore, in the bottom panel of online Appendix Table 7, we re-estimate our main OLS model in unstacked data (i.e., by subgroup) rather than in an interacted model.

Discussion

Consistent with the theoretical underpinnings of personnel economics (Lazear, 2000; Holmström, 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 public-school system (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. On average across teachers, we find dismissal threat effects on subsequent teacher performance of 0.22 teacher-level SD, which is roughly two-thirds of the difference in performance between novice and veteran teachers in DCPS (0.35 SD). For White veterans, dismissal threat effects upwards of 0.6 SD are larger than most teacher-oriented interventions, including scaled up, one-on-one instructional coaching programs (Kraft, Blazar, and Hogan 2018).

A more meaningful benchmark is the increase in teacher performance necessary to improve student outcomes. A growing body of evidence indicates that a 1 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 (Blazar & Kraft, 2017). This implies that the effects on classroom teaching performance 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 & Wolpin, 2003).

Nevertheless, 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. In several instances, we find larger effects of incentives on easier tasks with greatest expectation of

success, which is intuitive, consistent with longstanding theory (Atkinson, 1957), and aligned to other lab-based literature on this topic (Locke & Latham, 2002; Garbers & Konradt, 2014; Weibel, Rost, & Osterloh, 2010). These analyses, while exploratory, provide insight that teachers respond to incentives in ways and in areas in which they expect to do well.

The primary contribution of our paper focuses on heterogeneous responses by race and experience, where we argue that there also are evident links to expectancy. Consistent with several lab-based experiments that integrate race/ethnicity saliency into the incentive scheme (Farzana, Li, & Ren, 2015; Hoff & Pandey, 2006), we find that Black novices did not respond either to dismissal threats or to salary incentives. Black novices also were the least likely to reap the benefits of the incentives and, thus, likely had the lowest expectations of success. Aligned to the prior literature, we interpret these patterns of evidence of a potential self-fulfilling prophecy. Black novices perceived a lower likelihood of improvement and therefore responded less to the incentives. Over time, new cohorts of Black novices have priors about their likelihood of success based on cohorts that came before them.

Mapping expectations of success to incentive responses generally holds across additional subgroups. Compared to Black novices, Black veterans were less likely to face a dismissal threat and more likely to receive a salary increase offer; they responded to both incentives. White novices potentially had some of the highest expectations of success within the IMPACT system. They were threatened with dismissal and offered a salary incentive at similar rates to Black novices. But, they also voluntarily left the district following a dismissal threat at substantially higher rates, potentially signaling greater (perceived) outside opportunities. White novices also opted into the salary incentive at much higher rates than Black (and White) veterans. Aligned to these patterns, we find that White novices threatened with dismissal improved the most. At the top end of the performance

distribution, White novices and Black novices were both quite likely to earn a salary increase following the offer, even though White novices declined in their performance and Black novices improved (at least on one teaching task). This suggests that the two groups responded differently to the evaluation systems' requirement to *maintain* (but not necessarily improve) their high performance in order to receive a salary increase.

Responses to incentives by race and experience that we observe in our data further echo broader community concerns about racial inequities within IMPACT. Following the initial round of dismissals in spring 2010—experienced primarily by Black novices—teachers and community members raised concerns of inequities at community meetings and through protests. In fall 2010, the mayor who helped initiate IMPACT 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. Although the lawsuit did not focus on teacher race, public commentary did. While publicly 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). These inequities and racial tensions are further discussed in DCPS's own, more recent equity review (DCPS, 2021). Even though our own analyses do not point to clear racial biases amongst evaluators, we do show that, as a whole IMPACT, tended to reward White teachers much more than Black teachers. The attenuated responses of Black teachers echo the broader community's concerns that White teachers were better set up for success.

Heterogeneous effects by teacher race and experience not only provide empirical support for the role of expectancy theory in incentive schemes, but also suggest that the overall vision of

performance-based job incentives for teachers is not working fully as intended. Teacher merit pay has been implemented in U.S. school systems for many decades (e.g., Moore-Johnson, 1984; Murnane & Cohen, 1985), and has expanded substantially over the past 15 years following a federal incentive program that encouraged states and school districts to attach performance metrics to teacher job decisions (Howell & Magazinnik, 2017; McGuinn, 2012). Teacher evaluation systems now constitute upwards of four-tenths of school system budgets and roughly $2.4 billion in annual expenditures across the U.S. (Bleiberg et al., 2023; Chambers et al., 2013). In DCPS, our back-of-the envelope calculations come to a conservative estimate of roughly $4,800 per teacher per year (in 2023 dollars), which includes the financial incentives and the cost of performance monitoring.[7]

The rise of teacher evaluation systems in the U.S. also coincides with a vast increase in scholarship—including several experiments—on the importance of Black and other teachers of color to students' educational outcomes (e.g., Blazar, 2024; Gershenson et al., 2022; Redding, 2019), and growing policy attention for diversifying the teacher workforce (DeRamus-Byers, 2021; Education Commission of the States, 2019). Our findings suggest that these two policy goals may be difficult to reconcile. How can supply meet demand if Black teachers—and Black novices in particular—are less likely to respond to incentives than their White colleagues and, thus, more likely to be dismissed? Equally important, how can Black teachers feel supported and welcomed in education systems and larger labor markets when expectations of success in incentive schemes are low?

[7] We observe the exact amount of one-time bonuses, which comes to an average of roughly $3,000 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 $700 per teacher per year. We estimate an additional $1,000 per teacher per year for performance monitoring, primarily through classroom observations. Teachers were observed an average of four times per year; we assume 1.5 hours to prepare for, conduct, and debrief observations, and $100 per hour. This is a conservative estimate because it does not include costs associated with the hiring process because they are highly variable depending on the qualifications of outgoing/incoming teachers.

DCPS is somewhat of a unique case for exploring and unraveling these policy tensions and striving for greater equity, and some readers may be concerned that our findings cannot generalize to other contexts. The teacher evaluation incentives are particularly high-stakes in DCPS relative to other contexts (Bleiberg et al., 2023). Further, DCPS has one of the largest shares of Black teachers among all U.S. cities (Lindsey, Blom, & Tilsley, 2017). That said, attenuated incentive effects for Black teachers may be *more* severe in other areas where being Black makes one an "outsider" not just at the national level but also within one's own school and district.

Conclusion

What can we conclude about job-embedded performance incentives for teachers, given heterogeneous responses that extract social costs in addition to monetary ones? From an economic perspective—upon which school-based accountability, evaluation, and incentive schemes are built—a strict reading of Akerloff and Kranton's (2000) theory on identity economics and its parallels to expectancy theory (Lloyd & Mertens, 2018) suggests that Black teachers—and especially Black novices—need larger incentives 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 is not practical because it is unlawful, at least in the U.S.[8] An alternative may be to alter eligibility for incentives based on school characteristics that often correlate with teacher race, such as neighborhood and student income. DCPS leadership took this approach when they limited eligibility for salary incentives to teachers who worked in high-poverty schools, where Black teachers (and Black students) were overrepresented. At the same time, both approaches are indirect responses for building up expectations of success of Black teachers and sidestep broader concerns about accountability and evaluation systems being highly

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

racialized at their core (Au, 2016; Darling-Hammond, 2007). Teacher evaluation and incentives cannot be another example of reinforcing "outsider" identity and creating barriers to success for Black and other individuals of color.

References

Achinstein, B., & Ogawa, R. T. (2012). New teachers of color and culturally responsive teaching in an era of educational accountability: Caught in a double bind. *Journal of Educational Change*, *13*, 1-39.

Adnot, M., Dee, T., Katz, V., & Wyckoff, J. (2017). Teacher turnover, teacher quality, and student achievement in DCPS. *Educational Evaluation and Policy Analysis, 39*(1), 54-76.

Akerlof, G. A., & Kranton, R. E. (2000). Economics and identity. *The Quarterly Journal of Economics, 115*(3), 715-753.

Akerlof, G. A., & Kranton, R. E. (2005). Identity and the economics of organizations. *Journal of Economic Perspectives, 19*(1), 9-32.

Atkinson, J. W. (1957). Motivational determinants of risk-taking behavior. *Psychological Review, 64*(6p1), 359-372.

Au, W. (2016). Meritocracy 2.0: High-stakes, standardized testing as a racial project of neoliberal multiculturalism. *Educational Policy*, *30*(1), 39-62.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289-300.

Betts, J. R., Costrell, R. M., Walberg, H. J., Phillips, M., & Chin, T. (2001). Incentives and equity under standards-based reform. *Brookings Papers on Education Policy*, (4), 9-74.

Blazar, D. (Online 2024). Why Black teachers matter. *Educational Researcher*.

Blazar, D., & Kraft, M. A. (2017). Teacher and teaching effects on students' attitudes and behaviors. *Educational Evaluation and Policy Analysis, 39(*1), 146-170.

Bleiberg, J., Brunner, E., Harbatkin, E., Kraft, M. A., & Springer, M. G. (2023). *Taking teacher evaluation to scale: The effect of state reforms on achievement and attainment* (No. w30995). National Bureau of Economic Research.

Borjas, G. (2020). *Labor Economics* (8th ed.). McGraw Hill.

Bureau of Labor Statistics. (2020). Industries with largest employment. United States Department of Labor. https://www.bls.gov/emp/tables/industries-largest-employment.htm

Calonico, S., Cattaneo, M. D., & Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, *82*(6), 2295-2326.

Campbell, D. T. (1969). Reforms as experiments. *American Psychologist, 24(*4), 409-429.

Campbell, C. M., Parker, C., Shand, R., Kelly-Massound, A., Fashola, T., & Blanc, J. (2021). Perspectives on DCPS IMPACT teacher evaluation system: Findings from teachers and school leaders.

Campbell, S. L. (2023). Ratings in black and white: A quantcrit examination of race and gender in teacher evaluation reform. *Race Ethnicity and Education, 26*(7), 815-833.

Campbell, S. L., & Ronfeldt, M. (2018). Observational evaluation of teachers: Measuring more than we bargained for?. *American Educational Research Journal*, *55*(6), 1233-1267.

Cardoza, K. (2011, June 30). WTU Pres. calls teacher evaluations 'racist' ahead of ratings release. *WAMU American University Radio.* https://wamu.org/story/11/06/30/wtu-pres-calls-teacher-evaluations-racist-ahead-of-ratings-release/

Cattaneo, M. D., Jansson, M., & Ma, X. (2020). Simple local polynomial density estimators. *Journal of the American Statistical Association*, *115*(531), 1449-1455.

Cattaneo, M. D., & Titiunik, R. (2022). Regression discontinuity designs. *Annual Review of Economics*, *14*, 821-851.

Chambers, J., Brodziak de los Reyes, I., & O'Neil, C. (2013). How much are districts spending to implement teacher evaluation systems? Washington, D.C.: RAND Corporation.
Danielson, C. (2011). *Enhancing professional practice: A framework for teaching* (2nd ed.). Alexandria, VA: Association for Supervision and Curriculum Development.
Darling-Hammond, L. (2007). Race, inequality and educational accountability: The irony of 'No Child Left Behind'. *Race Ethnicity and Education*, *10*(3), 245-260.
Dee, T. S., & Wyckoff, J. (2015). Incentives, selection, and teacher performance: Evidence from IMPACT. *Journal of Policy Analysis and Management, 34(*2), 267-297.
Dee, T. S., James, J., & Wyckoff, J. (2021). Is effective teacher evaluation sustainable? Evidence from District of Columbia Public Schools. *Education Finance and Policy, 16(*2), 313-346.
DeRamus-Byers, R. (2021, July 12). Grow your own and teacher diversity in state legislative sessions: What we can learn from successfully passed bills. *New America*. https://www.newamerica.org/education-policy/edcentral/grow-your-own-and-teacher-diversity-in-state-legislative-sessions/
District of Columbia Public Schools. (2010). *IMPACT guidebook 2010-2011*.
District of Columbia Public Schools. (2021). *Equity review memo*.
District of Columbia Public Schools. (2021). *Initial set of evolutions to IMPACT: SY 21-22*.
District of Columbia Public Schools. (2022). *Evolutions to IMPACT: SY 22-23*.
Dong, N., & Maynard, R. (2013). PowerUp!: A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental design studies. *Journal of Research on Educational Effectiveness*, *6*(1), 24-67.
Eccles, J. S., & Wigfield, A. (2020). From expectancy-value theory to situated expectancy-value theory: A developmental, social cognitive, and sociocultural perspective on motivation. *Contemporary Educational Psychology, 61,* 101859.
Education Commission of the States. (2019, October 21). State information request: Diversifying the teacher workforce.
Ehrenberg, R. G., & Smith, R. S. (2016). *Modern labor economics: Theory and public policy* (12th ed.). Routledge.
Afridi, F., Li, S. X., & Ren, Y. (2015). Social identity and inequality: The impact of China's hukou system. *Journal of Public Economics, 123*, 17-29.
Figlio, D. N. (2002). Can public schools buy better-qualified teachers? *ILR Review, 55*(4), 686-699.
Fryer, R. G., Jr. (2017). The production of human capital in developed countries: Evidence from 196 randomized field experiments. In A. V. Banerjee & E. Duflo (Eds.), *Handbook of economic field experiments* (Vol. 2, pp. 95-322). North-Holland.
Garbers, Y., & Konradt, U. (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.
Gershenson, S., Hart, C. M. D., Hyman, J., Lindsay, C., & Papageorge, N. W. (2022). The long-run impacts of same-race teachers. *American Economic Journal: Economic Policy, 14*(4), 300-342.
Gneezy, U., & Rustichini, A. (2000). Pay enough or don't pay at all. *The Quarterly Journal of Economics, 115*(3), 791-810.
Hamre, B. K., Pianta, R. C., Downer, J. T., DeCoster, J., Mashburn, A. J., Jones, S. M., Brown, J. L., Cappella, E., Atkins, M., & Rivers, S. E. (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, E. A. (1979). Conceptual and empirical issues in the estimation of educational production functions. Journal of Human Resources, 14(3), 351-388.
Hanushek, E. A. (2007). The single salary schedule and other issues of teacher pay. *Peabody Journal of Education, 82*(4), 574-586.
Heilig, J. V., & Darling-Hammond, L. (2008). Accountability Texas-style: The progress and learning of urban minority students in a high-stakes testing context. *Educational Evaluation and Policy Analysis*, *30*(2), 75-110.
Hill, H. C., Blazar, D., & Lynch, K. (2015). Resources for teaching: Examining personal and institutional predictors of high-quality instruction. *AERA Open, 1*(4), 2332858415617703.
Hoff, K., & Pandey, P. (2006). Discrimination, social identity, and durable inequalities. *American Economic Review, 96*(2), 206-211.
Holmström, B. (1979). Moral hazard and observability. *The Bell Journal of Economics, 10*(1), 74-91.
Holmström, B., & Milgrom, P. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, and Organization, 7*(Special Issue), 24-52.
Howell, W. G., & Magazinnik, A. (2017). Presidential prescriptions for state policy: Obama's Race to the Top initiative. *Journal of Policy Analysis and Management, 36*(3), 502-531.
Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics, 142*(2), 615-635.
Irwin, V., Zhang, J., Wang, X., Hein, S., Wang, K., Roberts, A., York, C., Barmer, A., Bullock Mann, F., Dilig, R., & Parker, S. (2021). *Report on the condition of education 2021* (NCES 2021-144). National Center for Education Statistics.
Jacob, B. A., & Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics, 118*(3), 843-877.
Jiang, J. Y., & Sporte, S. E. (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, T. J., & Staiger, D. O. (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, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011). Identifying effective classroom practices using student achievement data. *Journal of Human Resources, 46*(3), 587-613.
Kraft, M. A., Blazar, D., & Hogan, D. (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, M. A., & Gilmour, A. F. (2017). Revisiting the widget effect: Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Researcher, 46*(5), 234-249.
Lazear, E. P. (2000). The power of incentives. *American Economic Review, 90*(2), 410-414.
Lazear, E. P. (2003). Teacher incentives. *Swedish Economic Policy Review, 10*(2), 179-214.
Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature, 4*8(2), 281-355.
Levitt, S. D., & List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives, 21*(2), 153-174.

Lindsay, C. A., Blom, E., & Tilsley, A. (2017). *Diversifying the classroom: Examining the teacher pipeline*. Urban Institute.

Lloyd, R., & Mertens, D. (2018). Expecting more out of expectancy theory: History urges inclusion of the social context. *International Management Review, 14*(1), 28-43.

Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist, 57*(9), 705-717.

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

McGuinn, P. (2012). Stimulating reform: Race to the Top, competitive grants and the Obama education agenda. *Educational Policy, 26*(1), 136-159.

Moore Johnson, S. (1984). Merit pay for teachers: A poor prescription for reform. *Harvard Educational Review, 54*(2), 175-186.

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

Omi, M., & Winant, H. (2014). *Racial formation in the United States*. Routledge.

Papay, J. P., & Kraft, M. A. (2015). Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement. *Journal of Public Economics, 130*, 105-119.

Pham, L. D., Nguyen, T. D., & Springer, M. G. (2021). Teacher merit pay: A meta-analysis. *American Educational Research Journal, 58*(3), 527-566.

Phipps, A. R., & Wiseman, E. A. (2021). Enacting the rubric: Teacher improvements in windows of high-stakes observation. *Education Finance and Policy, 16*(2), 283-312.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38*(2), 109-119.

Prendergast, C. (1999). The provision of incentives in firms. *Journal of Economic Literature, 37*(1), 7-63.

Putnam, H., Ross, E., & Walsh, K. (2018). *Making a difference: Six places where teacher evaluation systems are getting results*. National Council on Teacher Quality.

Quick, K. (2015). The unfair effects of IMPACT on teachers with the toughest jobs. *The Century Foundation*.

Reardon, S. F., & Robinson, J. P. (2012). Regression discontinuity designs with multiple rating-score variables. *Journal of Research on Educational Effectiveness, 5*(1), 83-104.

Redding, C. (2019). A teacher like me: A review of the effect of student–teacher racial/ethnic matching on teacher perceptions of students and student academic and behavioral outcomes. *Review of Educational Research, 89*(4), 499-535.

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

Sartain, L., & Steinberg, M. P. (2016). Teachers' labor market responses to performance evaluation reform: Experimental evidence from Chicago public schools. *Journal of Human Resources, 51*(3), 615-655.

Steele, C. M., Spencer, S. J., & Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. In *Advances in experimental social psychology* (Vol. 34, pp. 379-440). Academic Press.

Steinberg, M. P., & Sartain, L. (2021). What explains the race gap in teacher performance ratings? Evidence from Chicago Public Schools. *Educational Evaluation and Policy Analysis, 43*(1), 60-82.

Steinberg, M. P., & Sartain, L. (2015). Does teacher evaluation improve school performance? Experimental evidence from Chicago's Excellence in Teaching project. *Education Finance and Policy, 10*(4), 535-572.

Taylor, E. S., & Tyler, J. H. (2012). The effect of evaluation on teacher performance. *American Economic Review, 102*(7), 3628-3651.

Todd, P. E., & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal, 113*(485), F3-F33.

Tuma, A. P., Hamilton, L. S., & Tsai, T. (2018). *How Do Teachers Perceive Feedback and Evaluation Systems?: Findings from the American Teacher Panel*. RAND.

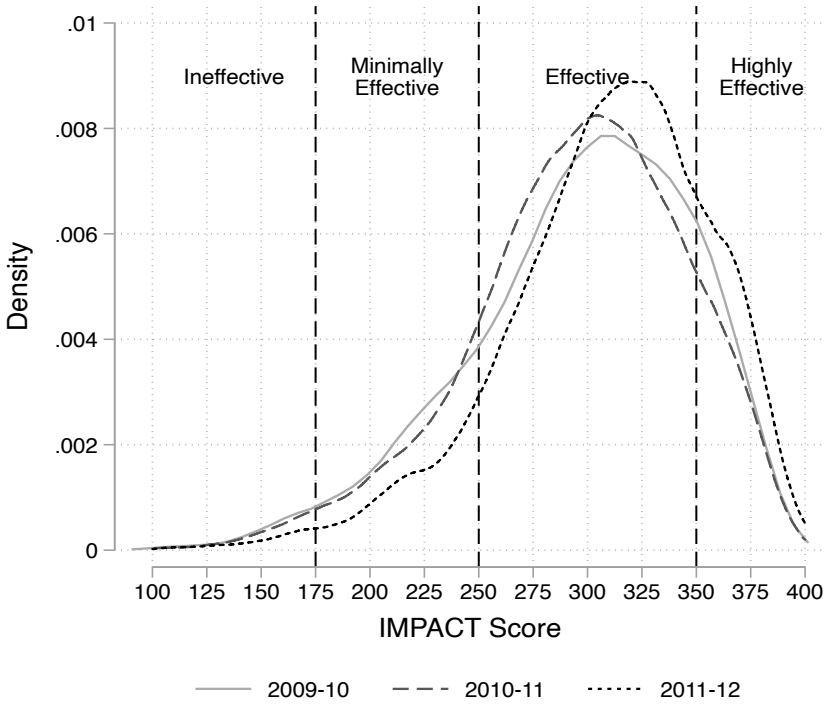
Weibel, A., Rost, K., & Osterloh, M. (2010). Pay for performance in the public sector—Benefits and (hidden) costs. *Journal of Public Administration Research and Theory, 20*(2), 387-412.

Whitmire, R. (2011). *The Bee Eater: Michelle Rhee takes on the nation's worst school district*. San Francisco: John Wiley & Sons.

Wiggins, G., & McTighe, J. (2005). *Understanding by design*. Alexandria: Association for Supervision and Curriculum Development.

Wong, V. C., Steiner, P. M., & Cook, T. D. (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



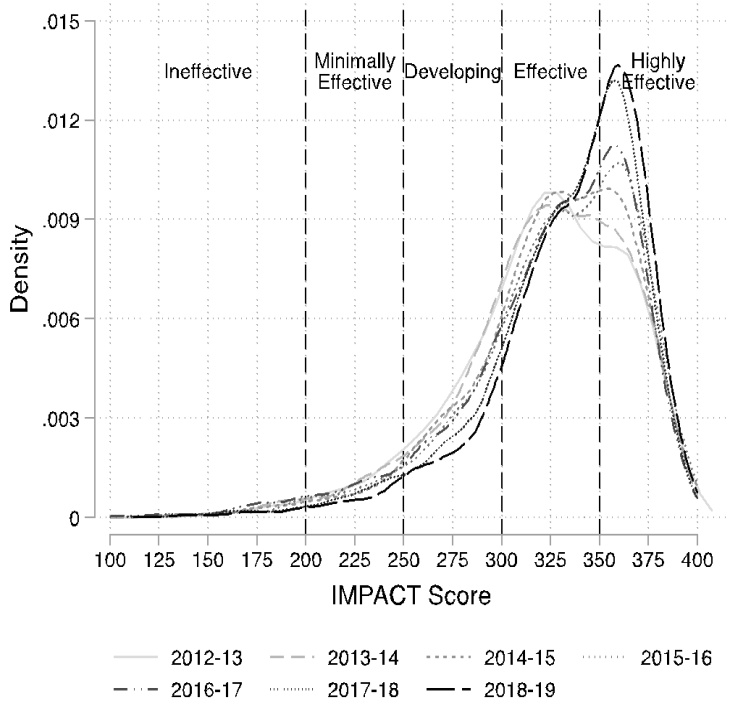


Figure 1: Distribution of Summative IMPACT Evaluation Scores and Associated Ratings.

Figure 2: Incentive Consequences and Rewards, by Teacher Race and Experience.

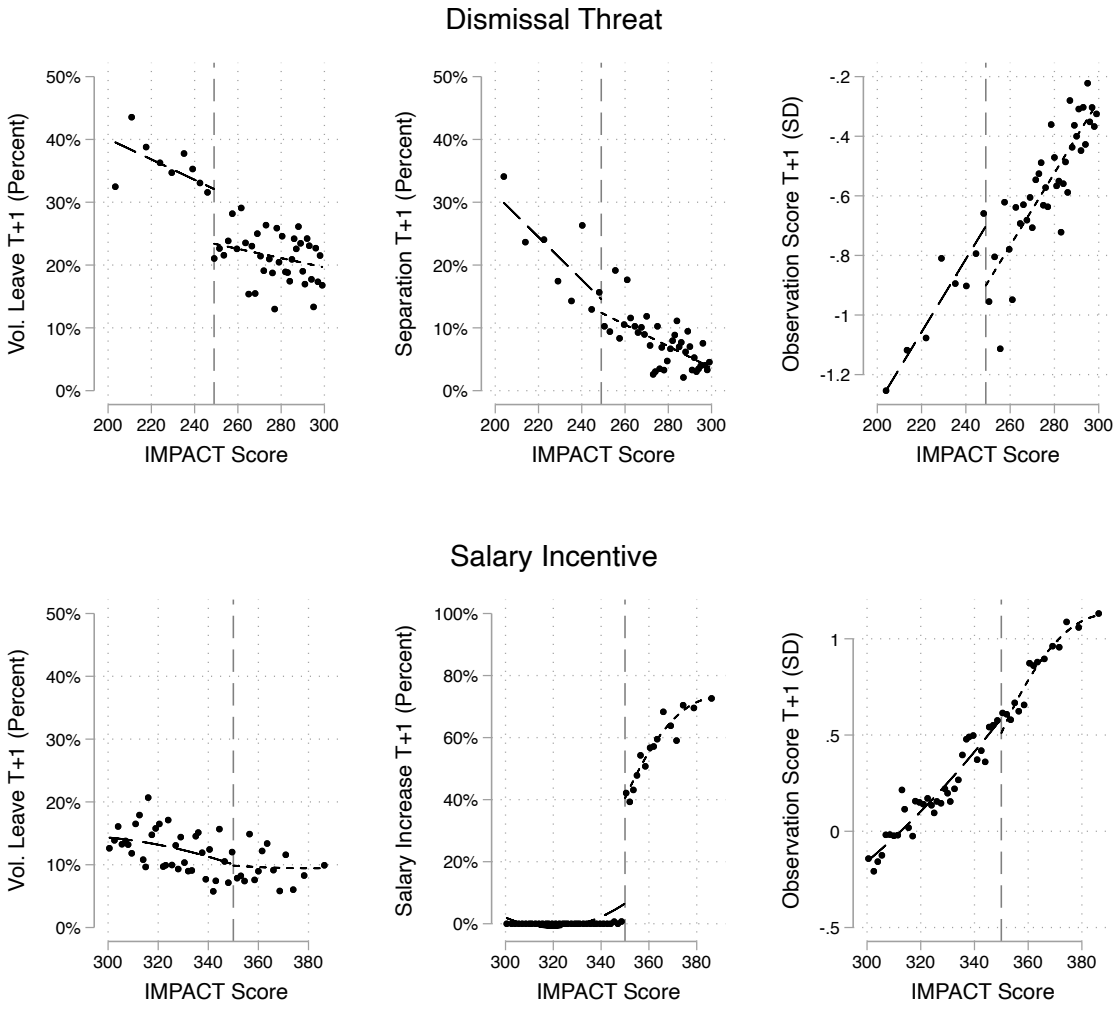


Figure 3: RD Impacts on Subset of Outcomes, on Average Across Teachers, Using 50 Bins of Equal Length.

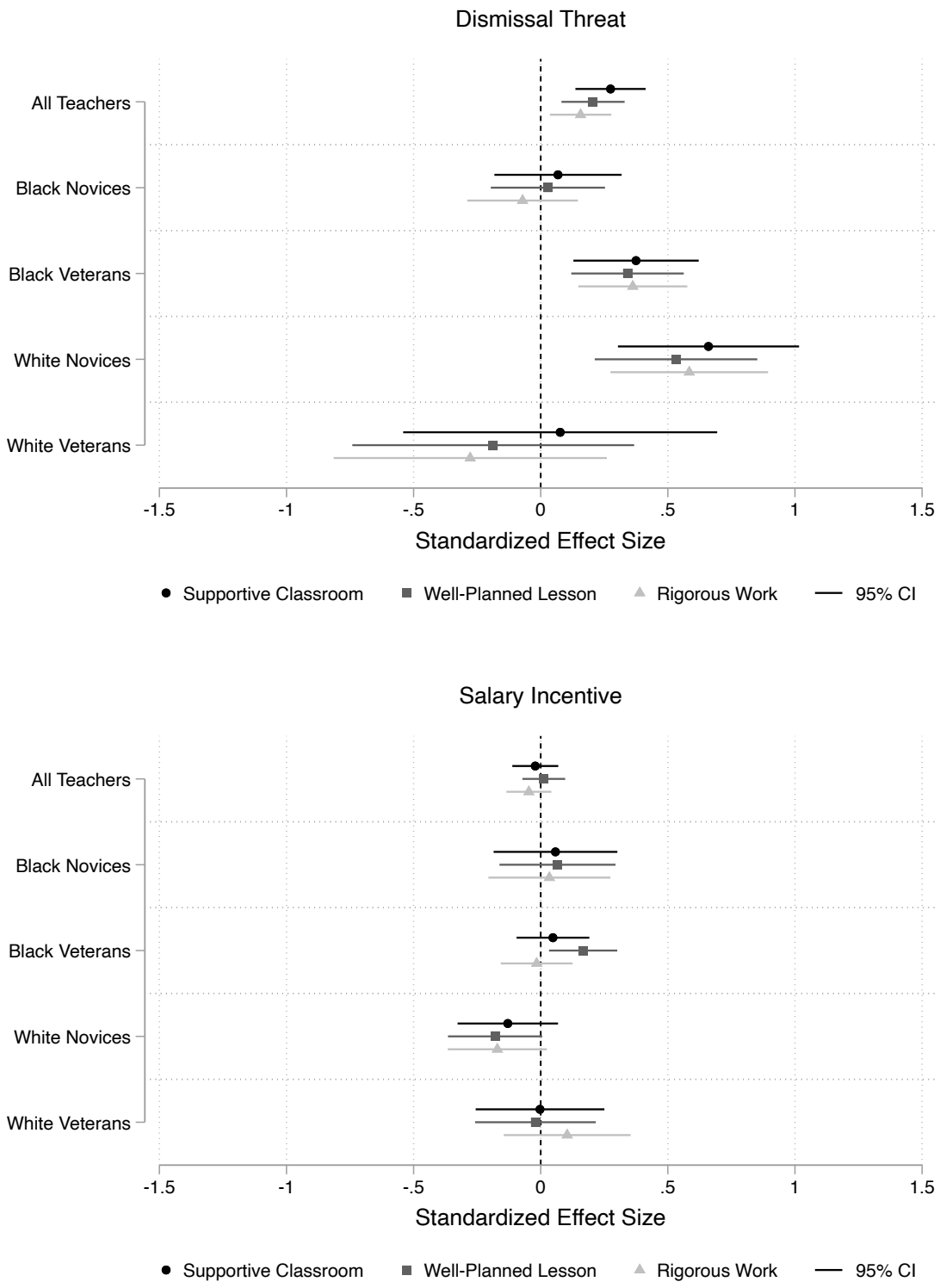


Figure 4. RD Impact Estimates and 95% Confidence Intervals, by Race*Experience and Teaching Task.

Tables

Table 1. Sample Descriptive Statistics

	All Teachers	Dismissal Threat Sample	Salary Incentive Sample
Asian	0.04	0.03	0.04
Black	0.50	0.57	0.54
Hispanic	0.05	0.05	0.04
White	0.32	0.23	0.32
Race/Ethnicity Missing	0.09	0.12	0.07
Female	0.74	0.69	0.76
Male	0.24	0.28	0.22
Gender Missing	0.02	0.03	0.02
Teaching Exp.: year 1	0.18	0.33	0.13
Teaching Exp.: years 2 to 4	0.30	0.26	0.29
Teaching Exp.: years 5 to 9	0.19	0.12	0.20
Teaching Exp.: years 10 to 19	0.17	0.13	0.19
Teaching Exp. 20 years or more	0.15	0.15	0.18
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	0.65	0.63	0.66
Teach Special Ed.	0.16	0.14	0.18
Teach English Language Learners	0.03	0.03	0.05
High-Poverty School	0.78	0.89	0.82
Low-Poverty School	0.22	0.11	0.18
Early Childhood	0.17	0.17	0.19
Elementary School	0.48	0.42	0.47
Middle School	0.25	0.29	0.25
High School	0.10	0.13	0.09
Dismissal Threat Sample	0.19	1.00	0.00
Dismissal Threat Offer	0.04	0.20	0.00
Salary Incentive Sample	0.23	0.00	1.00
Salary Incentive Offer	0.10	0.00	0.29
Has Outcome Data in Year T+1	0.80	0.74	0.86
Teachers	8,907	4,235	3,927
Teacher-Year Observations	30,726	5,824	7,108

Table 2. Differences in Summary Observation Score (Lesson Level) Across Race/Ethnicity Groups

	(1)	(2)	(3)	(4)	(5)	Prop. of Lessons with Same-Race/ Ethnicity Rater
Asian	0.097*	0.133***	0.066~	0.101**		
	(0.040)	(0.039)	(0.037)	(0.037)		
Asian*Rater Match				-0.134	-0.090	0.01
				(0.179)	(0.155)	
Black*Rater Match				0.058***	0.074***	0.60
				(0.012)	(0.010)	
Hispanic	0.052	0.097**	-0.032	0.014		
	(0.035)	(0.035)	(0.036)	(0.037)		
Hispanic*Rater Match				-0.125*	-0.038	0.13
				(0.064)	(0.048)	
White	0.299***	0.343***	0.181***	0.217***		
	(0.016)	(0.017)	(0.017)	(0.019)		
White*Rater Match				-0.010	0.002	0.36
				(0.014)	(0.012)	
Observations	83,653	83,653	83,653	83,653	83,653	
Teacher Experience		X	X	X	X	
School Fixed Effects			X	X		
Teacher Fixed Effects					X	

Notes: Estimates in each column come from the same model that regresses teachers' lesson-level summary observation score on race/ethnicity dummies. In models (1) through (4), Black is the left-out/reference category. All models include fixed effects for school year, an indicator for whether or not the lesson was scored by a school leader versus master educator, and the order of the observation in the school year. Teachers with missing race/ethnicity information are excluded from the analysis.

Table 3. Balance and Sorting Tests across the Eligibility Threshold

	Dismissal Threat		Salary Incentive	
	Yr. T	Yr. T+1	Yr. T	Yr. T+1
Female	0.012	-0.003	-0.054	-0.070~
	(0.055)	(0.065)	(0.037)	(0.039)
Asian	-0.039*	-0.049*	0.005	0.002
	(0.017)	(0.022)	(0.018)	(0.020)
Black	0.032	0.045	-0.051	-0.025
	(0.058)	(0.068)	(0.045)	(0.048)
Hispanic	-0.011	0.030	0.030~	0.031
	(0.025)	(0.032)	(0.017)	(0.019)
White	-0.023	-0.034	0.000	0.006
	(0.048)	(0.058)	(0.042)	(0.045)
Novice	-0.056	-0.051	0.072~	0.063
	(0.056)	(0.068)	(0.043)	(0.046)
Teach Gen. Ed., Tested Grades/Subjects	-0.039	0.009	-0.031	-0.042
	(0.050)	(0.060)	(0.026)	(0.028)
Teach Gen. Ed., Non-Tested Grades/Subj.	0.056	0.038	0.002	-0.001
	(0.056)	(0.067)	(0.042)	(0.044)
Teach Special Education	-0.012	-0.046	-0.003	0.018
	(0.039)	(0.043)	(0.033)	(0.034)
Teach English Language Learners	-0.005	-0.001	0.032~	0.026
	(0.014)	(0.016)	(0.019)	(0.020)
High-Poverty School	-0.041	-0.084*	0.034	0.044
	(0.030)	(0.040)	(0.032)	(0.035)
Early Childhood	0.055	0.043	0.071~	0.073~
	(0.043)	(0.049)	(0.037)	(0.040)
Elementary School	-0.007	0.007	0.004	-0.002
	(0.057)	(0.069)	(0.045)	(0.048)
Middle School	-0.081	-0.115~	-0.050	-0.034
	(0.052)	(0.061)	(0.038)	(0.040)
High School	0.033	0.064	-0.026	-0.037
	(0.041)	(0.048)	(0.026)	(0.027)
Observations	5,824	4,291	7,101	6,128
P-Value on Joint Test of Significance	0.447	0.251	0.208	0.171
P-Value on Density Test	0.694	0.118	0.331	0.157

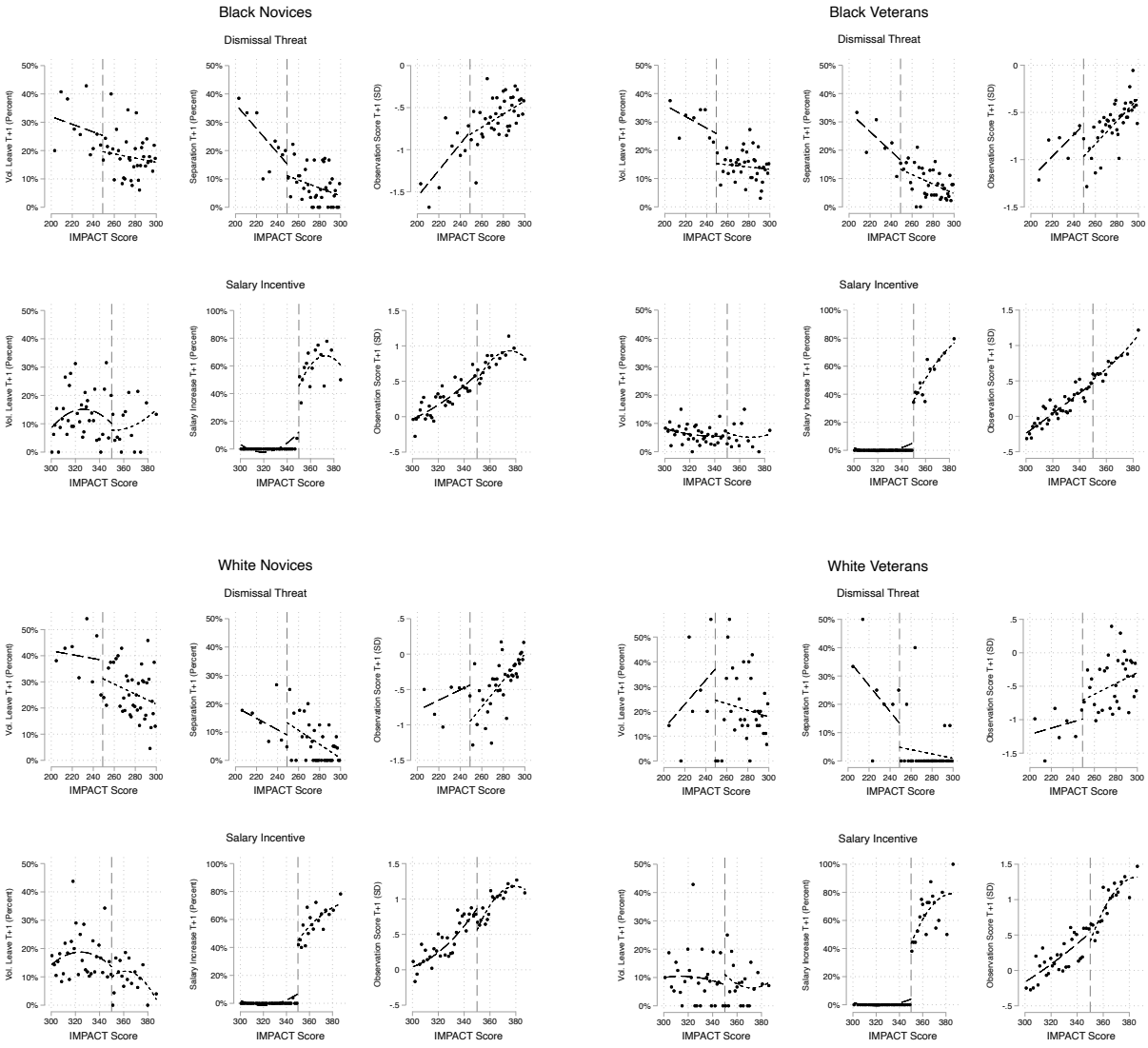
Notes: Estimates in each cell are from separate models that regress the teacher or school covariate listed on a dummy indicator for dismissal threat or base salary increase offer, a cubic function of distance from threshold that determines eligibility for the incentive (where the function varies on either side of the threshold), and year fixed effects. Heteroskedasticity-robust standard errors clustered at the teacher level in parentheses. *** p<0.001, ** p<0.01, * p<0.05, ~ p<0.1.

Table 4. RD Estimates of Differential Responses to Evaluation Incentives by Race*Experience

	Dismissal Threat					Salary Incentive				
	N Yr. T	N Yr. T+1	Voluntary Leave Yr. T	Separation Yr. T+1	Obs. Score T+1	N Yr. T	N Yr. T+1	Voluntary Leave Yr. T	Salary Increase Yr. T+1	Obs. Score T+1
All Teachers	5,824	4,291	0.073**	0.033	0.217***	7,101	6,128	-0.020	0.375***	-0.030
			(0.024)	(0.023)	(0.064)			(0.017)	(0.025)	(0.039)
Black Novices	1547	1197	0.053	0.060	0.006	951	816	-0.035	0.438***	0.030
			(0.044)	(0.042)	(0.116)			(0.049)	(0.068)	(0.095)
Black Veterans	1763	1416	0.115**	0.039	0.380***	2867	2639	0.013	0.336***	0.066
			(0.042)	(0.042)	(0.115)			(0.022)	(0.042)	(0.060)
White Novices	1027	718	0.029	-0.042	0.606***	1421	1175	-0.066	0.412***	-0.176*
			(0.065)	(0.052)	(0.153)			(0.040)	(0.054)	(0.089)
White Veterans	331	242	0.033	0.150~	-0.178	805	725	-0.004	0.390***	-0.020
			(0.111)	(0.089)	(0.275)			(0.045)	(0.070)	(0.123)
P-Values on Differential Effects between Race/Experience Groups										
Black Novices v. Veterans			0.309	0.721	**0.021**			0.369	0.203	0.749
White Novices v. Veterans			0.973	**0.064**	**0.013**			0.300	0.800	0.306
Black v. White Novices			0.759	0.130	**0.002**			0.623	0.764	0.113
Black v. White Veterans			0.491	0.261	**0.061**			0.743	0.509	0.530
Black Novices v. White Veterans			0.869	0.362	0.537			0.637	0.619	0.748
Black Veterans v. White Novices			0.265	0.233	0.238			**0.086**	0.268	**0.024**

Notes: All regression models control for year fixed effects, observable teacher and school characteristics (see Table 1), and a function of distance from the threshold that determines eligibility for the incentive (linear for dismissal threat and quadratic for salary incentive). Heteroskedasticity-robust standard errors clustered at the teacher level in parentheses. *** p<0.001, ** p<0.01, * p<0.05, ~ p<0.1 on estimates of treatment effects. For differential effects between groups, exact p-values are reported and those below 0.1 are in bold.

Appendix



Appendix Figure 1. RD Impacts of Dismissal Threat, by Subgroup, Using 50 Bins of Equal Length.

Appendix Table 1. IMPACT Design Features and Changes Over Time

	2009-10 to 2011-12	2012-13 to 2013-14	2014-15 to 2015-16	2016-17 and After
Performance Monitoring		*Observation Instrument*		
		Teaching and Learning Framework (TLF) with 9 tasks		Essential Practices (EP) with 5 tasks
		Number of Observations		
		Up to 5 observations: 3 from school leader and 2 from master educator		Up to 3 observations, all by school leader
		Performance Bands		
	4 performance bands: Ineffective (I), Minimally Effective (ME), Effective (E), Highly Effective (HE)	5 performance bands, splitting E in two to include Developing (D)		
		Percent of IMPACT Score Based on Observations		
	35% for general education teachers in tested grade/subject (Group 1), 75% for general education teachers in non-tested grade/subject (Group 2) and teachers of English language learners (Group 4), and 65% for special education teachers (Group 3)	40% for Group 1 and same for other 3 groups	75% for Group 1 and same for other 3 groups	30% for Group 1, 65% to 75% for Grpup 2, and same for other 2 groups
Dismissal Threat	Separation after 1 I rating, or 2 consecutive ME ratings	Separation after 1 I rating, 2 consecutive ME ratings, 1 D followed by 1 ME rating, or 3 consecutive ratings below E		
Salary Incentive	Base pay increase after 2 consecutive HE ratings: MA degree band plus 3 service credits for low-poverty schools or 5 credits for high-poverty schools	Base pay increase only available to teachers in high-poverty schools and based on career ladder: Advanced teachers (i.e., 1 HE or 2 consecutive E ratings) received 2 service credits, Distinguished teachers (i.e., 2 consecutive HE rating, after getting to Advanced) receive MA degree band plus 5 service credits, Expert teacher (i.e., 2 consecutive HE rating, after getting to Distinguished) receive PhD degree band plus 5 service credits		

Appendix Table 2. Intraclass Correlations as Sorting Tests of Raters to Teachers

Baseline Teacher Characteristics	School Leaders: Within School and Year (all years)	Master Educators (2009-2010 to 2015-16)	
		Within School and Year	Within Year, Across Schools
	Panel A: Conditional on Leave-Out, School-Year Average		
Summary Observation Score T-1	0.000	0.010	0.010
Asian	0.000	0.014	0.014
Black	0.003	0.036	0.032
Hispanic	0.032	0.077	0.077
White	0.002	0.012	0.010
Novice	0.000	0.018	0.015
	Panel B: Unconditional		
Summary Observation Score T-1	0.028	0.011	0.018
Asian	0.002	0.015	0.019
Black	0.014	0.038	0.077
Hispanic	0.030	0.077	0.139
White	0.011	0.014	0.020
Novice	0.018	0.022	0.030

Notes: Estimates are intraclass correlations (ICCs) that estimate how much of the variation in a given baseline teacher characteristic lies within versus between raters. ICCs are calculated from multi-level models that all condition on year fixed effects; some models also condition on school fixed effects. In Panel A, models further condition on a school-year, leave-out-average of the baseline teacher characteristic; in Panel B, these covariates are removed. *** $p<0.001$, ** $p<0.01$, * $p<0.05$, ~ $p<0.1$.

Appendix Table 3. Incentive Rollout and Take-up

		Dismissal Threat				Salary Incentive		
	N	Immediate Separation	Dismissal Threat	Voluntary Leave if Threatened	Separation after Threat if Stayed	Salary Offer	Opt-In if Offered	Salary Increase if Offered and Opted In
All Teachers	30,726	0.017	0.038	0.360	0.212	0.097	0.605	0.673
Black Teachers	15,467	0.018	0.041	0.296	0.234	0.089	0.556	0.693
White Teachers	9,974	0.007	0.022	0.372	0.157	0.116	0.685	0.643
Novice Teachers	15,055	0.024	0.051	0.383	0.204	0.072	0.705	0.663
Veteran Teachers	15,671	0.011	0.025	0.314	0.227	0.120	0.537	0.679
Black Novices	5,533	0.026	0.061	0.287	0.238	0.056	0.759	0.716
Black Veterans	9,895	0.013	0.031	0.306	0.230	0.108	0.510	0.686
White Novices	5,906	0.007	0.029	0.398	0.130	0.097	0.763	0.633
White Veterans	4,019	0.003	0.012	0.277	0.235	0.145	0.580	0.653

Appendix Table 4. Crosswalk between Teaching Tasks across Rubrics, and Descriptive Statistics

Constructs Created for Analysis				Teaching and Learning Framework: 2009-10 to 2015-16		Essential Practices: 2016-17 to Present	
Domain	Mean	Teacher-Year ICC	Rater ICC	Domain	Mean	Domain	Mean
Build a supportive classroom	3.44	0.65	0.10	Build a supportive, learning-focused classroom	3.40	Cultivate a responsive learning community	3.51
Lead well-planned and responsive lessons	3.23	0.75	0.12	Maximize instructional time	3.27	Lead a well-planned, purposeful learning experience	3.30
				Check for student understanding	3.22		
				Provide students multiple ways to move toward mastery	3.19		
				Lead well-organized objective-driven lessons	3.18		
				Explain content clearly	3.17	Respond to evidence of student learning	3.24
				Respond to student mis-understandings	3.07		
Engage students in rigorous and higher-level work	2.97	0.66	0.15	Engage students at all learning levels in rigorous work	2.95	Challenge students with rigorous content	3.19
				Develop higher-level understanding	2.72	Maximize student ownership of learning	3.12

Note: Following a generalizability framework, teacher-year ICCs are adjusted for the modal number of lessons per teacher.

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

	Dismissal Threat							Salary Incentive						
	N	SC	WL	RW	P-Values on Differential Effects between Tasks			N	SC	WL	RW	P-Values on Differential Effects between Tasks		
					SC v. WL	SC v. RW	WL v. RW					SC v. WL	SC v. RW	WL v. RW
All Teachers	4,291	0.275***	0.206**	0.157*	0.201	**0.054**	0.178	6,128	-0.021	0.012	-0.047	0.416	0.593	**0.088**
		(0.071)	(0.063)	(0.062)					(0.046)	(0.043)	(0.045)			
Black Novices	1197	0.068	0.028	-0.071	0.672	0.197	0.110	816	0.058	0.066	0.034	0.950	0.867	0.745
		(0.128)	(0.115)	(0.111)					(0.124)	(0.116)	(0.122)			
Black Veterans	1416	0.375**	0.341**	0.362***	0.732	0.909	0.759	2639	0.048	0.167*	-0.016	**0.057**	0.394	**0.001**
		(0.126)	(0.113)	(0.109)					(0.073)	(0.068)	(0.072)			
White Novices	718	0.660***	0.532**	0.584***	0.379	0.629	0.567	1175	-0.130	-0.180~	-0.171~	0.562	0.685	0.899
		(0.182)	(0.163)	(0.158)					(0.101)	(0.094)	(0.099)			
White Veterans	242	0.077	-0.187	-0.277	0.273	0.205	0.596	725	-0.003	-0.021	0.104	0.888	0.428	0.210
		(0.315)	(0.283)	(0.274)					(0.129)	(0.121)	(0.127)			

Notes: SC = Supportive Classroom; WL = Well-Planned Lesson; RW = Rigorous Work. All regression models control for year fixed effects, observable teacher and school characteristics (see Table 1), and a function of distance from the threshold that determines eligibility for the incentive (linear for dismissal threat and quadratic for salary incentive). Heteroskedasticity-robust standard errors clustered at the teacher level in parentheses. *** p<0.001, ** p<0.01, * p<0.05, ~ p<0.1 on estimates of treatment effects. For differential effects between tasks, exact p-values are reported and those below 0.1 are in bold.

Appendix Table 6. Alternative Specifications of RD Estimates of Differential Responses to Evaluation Incentives

	Dismissal Threat					Salary Incentive				
	N Yr. T	N Yr. T+1	Voluntary Leave Yr. T	Separation Yr. T+1	Obs. Score T+1	N Yr. T	N Yr. T+1	Voluntary Leave Yr. T	Salary Increase Yr. T+1	Obs. Score T+1
					Panel A: Preferred Functional Form, with School Fixed Effects					
All Teachers	5,824	4,291	0.074**	0.038~	0.199**	7,101	6,128	-0.026	0.372***	-0.026
			(0.024)	(0.023)	(0.065)			(0.017)	(0.025)	(0.038)
Black Novices	1547	1197	0.063	0.064	0.004	951	816	-0.040	0.446***	0.012
			(0.045)	(0.043)	(0.117)			(0.049)	(0.069)	(0.094)
Black Veterans	1763	1416	0.129**	0.043	0.351**	2867	2639	0.006	0.330***	0.052
			(0.042)	(0.043)	(0.114)			(0.022)	(0.043)	(0.061)
White Novices	1027	718	0.028	-0.033	0.637***	1421	1175	-0.071~	0.407***	-0.139
			(0.066)	(0.050)	(0.154)			(0.041)	(0.054)	(0.086)
White Veterans	313	242	0.038	0.155~	-0.200	805	725	-0.007	0.395***	-0.011
			(0.114)	(0.091)	(0.288)			(0.045)	(0.068)	(0.122)
					Panel B: Alternative Functional Form, without School Fixed Effects					
All Teachers	5,824	4,291	0.022	0.036	0.196*	7,101	6,128	-0.048~	0.363***	-0.050
			(0.034)	(0.034)	(0.091)			(0.026)	(0.031)	(0.060)
Black Novices	1547	1197	0.015	0.107	-0.005	951	816	-0.094	0.403***	-0.022
			(0.062)	(0.068)	(0.170)			(0.086)	(0.093)	(0.144)
Black Veterans	1763	1416	0.060	0.004	0.325~	2867	2639	-0.062~	0.374***	0.107
			(0.062)	(0.061)	(0.172)			(0.033)	(0.051)	(0.094)
White Novices	1027	718	-0.005	-0.070	0.634**	1421	1175	-0.072	0.399***	-0.212
			(0.091)	(0.057)	(0.202)			(0.061)	(0.068)	(0.130)
White Veterans	313	242	0.009	0.211	-0.076	805	725	0.023	0.286**	-0.179
			(0.157)	(0.172)	(0.457)			(0.067)	(0.088)	(0.186)

Notes: All regression models control for year fixed effects, observable teacher characteristics (see Table 1), and function of distance from the threshold that determines eligibility for the incentive. For dismissal threat analyses, preferred functional form is linear and alternative is quadratic; for salary incentive salary, preferred functional form is quadratic and alternative is cubic. In Panel B, models that exclude school fixed effects further include observable school characteristics. Heteroskedasticity-robust standard errors clustered at teacher level in parentheses. *** p<0.001, ** p<0.01, * p<0.05, ~ p<0.1

Appendix Table 7. RD Estimates of Differential Responses to Evaluation Incentives, by Bandwidth

	Dismissal Threat					Salary Incentive				
	N Yr. T	N Yr. T+1	Voluntary Leave Yr. T	Separation Yr. T+1	Obs. Score T+1	N Yr. T	N Yr. T+1	Voluntary Leave Yr. T	Salary Increase Yr. T+1	Obs. Score T+1
					Panel A: Bandwidth of 40					
All Teachers	4,320	3,123	0.076**	0.031	0.220**	6,109	5,290	-0.018	0.366***	-0.010
			(0.027)	(0.026)	(0.075)			(0.017)	(0.026)	(0.041)
Black Novices	1179	903	0.031	0.076	0.016	808	688	-0.003	0.398***	-0.007
			(0.050)	(0.049)	(0.135)			(0.053)	(0.071)	(0.100)
Black Veterans	1287	1011	0.134**	0.023	0.390**	2439	2255	0.000	0.342***	0.113~
			(0.048)	(0.048)	(0.134)			(0.021)	(0.043)	(0.060)
White Novices	745	510	0.015	-0.053	0.521**	1229	1016	-0.050	0.425***	-0.180~
			(0.074)	(0.059)	(0.184)			(0.044)	(0.054)	(0.093)
White Veterans	216	162	0.102	0.099	-0.054	733	659	0.001	0.370***	-0.024
			(0.126)	(0.100)	(0.310)			(0.048)	(0.072)	(0.128)
					Panel B: Bandwidth of 30					
All Teachers	3,072	2,205	0.059~	0.007	0.284**	4,946	4,315	-0.026	0.360***	-0.032
			(0.031)	(0.030)	(0.088)			(0.019)	(0.028)	(0.045)
Black Novices	851	652	0.017	0.041	0.076	640	558	-0.005	0.436***	0.068
			(0.057)	(0.056)	(0.155)			(0.061)	(0.077)	(0.108)
Black Veterans	901	703	0.124*	0.011	0.475**	1961	1820	-0.023	0.346***	0.089
			(0.055)	(0.055)	(0.156)			(0.024)	(0.046)	(0.067)
White Novices	522	346	-0.001	-0.082	0.762***	1013	840	-0.059	0.393***	-0.226*
			(0.086)	(0.068)	(0.220)			(0.048)	(0.059)	(0.102)
White Veterans	151	112	0.092	0.121	-0.303	614	552	0.021	0.325***	-0.063
			(0.143)	(0.113)	(0.352)			(0.054)	(0.078)	(0.140)
					Panel C: Bandwidth of 20					
All Teachers	1,940	1,363	0.070~	0.038	0.261*	3,479	3,049	-0.044~	0.354***	-0.056
			(0.038)	(0.037)	(0.110)			(0.023)	(0.032)	(0.053)
Black Novices	548	416	-0.014	0.103	0.092	431	380	-0.035	0.436***	-0.005
			(0.069)	(0.071)	(0.193)			(0.076)	(0.092)	(0.131)
Black Veterans	564	428	0.139*	-0.007	0.596**	1362	1263	-0.045	0.340***	0.069
			(0.068)	(0.068)	(0.200)			(0.027)	(0.052)	(0.080)
White Novices	313	200	0.039	-0.033	0.309	732	611	-0.091	0.380***	-0.234*
			(0.102)	(0.084)	(0.280)			(0.056)	(0.067)	(0.119)
White Veterans	87	63	0.351*	0.199	-0.269	449	407	-0.017	0.307***	-0.084
			(0.160)	(0.143)	(0.454)			(0.061)	(0.091)	(0.171)

Notes: All regression models control for year fixed effects, observable teacher and school characteristics (see Table 1), and a function of distance from the threshold that determines eligibility for the incentive (linear for dismissal threat and quadratic for salary incentive). Heteroskedasticity-robust standard errors clustered at the teacher level in parentheses. *** p<0.001, ** p<0.01, * p<0.05, ~ p<0.1

Appendix Table 8. Local Polynomial and Linear Regression RD Estimates of Differential Responses to Evaluation Incentives, in Unstacked Dataset

	Dismissal Threat					Salary Incentive				
	N Yr. T	N Yr. T+1	Voluntary Leave Yr. T	Separation Yr. T+1	Obs. Score T+1	N Yr. T	N Yr. T+1	Voluntary Leave Yr. T	Salary Increase Yr. T+1	Obs. Score T+1
Panel A: Local Polynomial Regression with Robust Bias-Corrected Confidence Intervals, Unstacked Subgroups										
All Teachers	5824	4,291	0.047	0.034	0.224*	7101	6,128	-0.048~	0.374***	-0.066
			(0.038)	(0.038)	(0.110)			(0.026)	(0.032)	(0.061)
Black Novices	1,547	1,205	0.005	0.104	0.016	951	816	-0.100	0.417***	-0.076
			(0.069)	(0.073)	(0.207)			(0.085)	(0.096)	(0.141)
Black Veterans	1,763	1,421	0.099	0.013	0.376~	2,867	2,639	-0.062~	0.382***	0.107
			(0.067)	(0.068)	(0.200)			(0.033)	(0.051)	(0.095)
White Novices	1,027	718	0.004	-0.099	0.605*	1,421	1,175	-0.076	0.406***	-0.236~
			(0.103)	(0.071)	(0.261)			(0.061)	(0.071)	(0.131)
White Veterans	313	242	0.183	0.158	-0.235	805	725	0.032	0.302***	-0.193
			(0.160)	(0.177)	(0.455)			(0.068)	(0.089)	(0.201)
Panel B: Linear Regression, Unstacked Subgroups										
Black Novices	1,547	1,205	0.054	0.063	0.009	951	816	-0.038	0.420***	0.017
			(0.044)	(0.043)	(0.117)			(0.051)	(0.068)	(0.094)
Black Veterans	1,763	1,421	0.111**	0.034	0.367**	2,867	2,639	0.014	0.341***	0.069
			(0.041)	(0.042)	(0.116)			(0.022)	(0.042)	(0.060)
White Novices	1,027	718	0.029	-0.043	0.601***	1,421	1,175	-0.073~	0.404***	-0.187*
			(0.065)	(0.052)	(0.155)			(0.040)	(0.052)	(0.089)
White Veterans	313	242	0.010	0.132	-0.128	805	725	0.000	0.394***	-0.010
			(0.115)	(0.085)	(0.265)			(0.047)	(0.070)	(0.125)

Notes: All regression models control for year fixed effects and observable teacher and school characteristics (see Table 1). In Panel A, effects of dismissal threats are estimated as a local linear regression; effects of salary incentives are estimated as a local quadratic. In Panel B, models also include a function of distance from the threshold that determines eligibility for the incentive (linear for dismissal threat and quadratic for salary incentive). Heteroskedasticity-robust standard clustered at the teacher level errors in parentheses. *** p<0.001, ** p<0.01, * p<0.05, ~ p<0.1