Teacher Evaluation Reform and the Exit of Low-Performing Teachers: The Role of Tenure and the Timing of Evaluation Ratings

LAUREN SARTAIN AND MATTHEW P. STEINBERG*

November 2020

Teacher quality is the most important school factor for improving student outcomes. Yet, personnel evaluation systems have historically failed to identify and remediate low-performing teachers. As part of a nationwide movement, Chicago Public Schools implemented a new evaluation system in 2012, which incorporated remediation and dismissal plans for low-rated teachers. In this paper, we employ a regression discontinuity (RD) design to identify the impact of high-stakes evaluation on the exit of low-performing teachers, as well as heterogeneous effects by tenure status. Though tenured teachers are, on average, significantly less likely to exit Chicago than equivalently rated non-tenured teachers, RD estimates indicate that receipt of a low evaluation rating increased the exit from Chicago of low-rated tenured teachers by 50 percent. This increase was driven by involuntary exit. We also find that the teacher labor supply available to replace low-rated teachers is of higher quality on multiple dimensions, and the labor supply is sufficient to support the removal of significantly more low-rated teachers These findings suggest that teacher evaluation systems can differentiate and remediate low-performance while revealing that the available teacher labor supply is sufficient to improve the quality of the teacher labor force, a primary goal of teacher evaluation reform.

* Sartain: University of North Carolina at Chapel Hill School of Education, 107 Peabody Hall, CB 3500, Chapel Hill, NC 27599 (email: lsartain@email.unc.edu) and University of Chicago Consortium on School Research, 1313 E. 60th Street, Chicago, IL 60637. Steinberg: George Mason University College of Education and Human Development, 4400 University Drive, Fairfax, VA 22030 (email: msteinb6@gmu.edu). Authors listed alphabetically and are equal contributors to this article. The authors thank the staff at Chicago Public Schools, particularly in the Talent Office, and UChicago Consortium for providing access to the data and helping us better understand the policy context. We also thank seminar participants at Brown University, the University of North Carolina at Chapel Hill, the Association for Public Policy and Management annual conference, and the Association for Education Finance and Policy annual conference. Any errors are ours.

Introduction

One of the most persistent and urgent problems facing education policymakers is the provision of highly effective teachers in all of our nation's classrooms. The increasing demand for high-quality teachers, particularly in urban public schools, has been well documented for at least three decades (National Commission on Excellence in Education, 1983; Ingersoll, 2001; Murnane & Steele, 2007). Indeed, of all school-level factors related to student learning and achievement, the student's teacher has consistently been shown to be the most important (Goldhaber, 2002; Rockoff, 2004; Rivkin, Hanushek, & Kain, 2005). Empirical evidence confirms that high-quality teaching is influential for student achievement, socio-emotional, and labor market outcomes (Aaronson, Barrow, & Sander, 2007; Chetty, Friedman, & Rockoff, 2014; Goldhaber, 2002; Jackson et al., In Press; Kraft, 2019; Rockoff, 2004; Rivkin, Hanushek, & Kain, 2005). At the same time, teachers vary dramatically in their ability to improve student performance (Rivkin et al., 2005; Aaronson et al., 2007). The strong causal relationship between teacher quality and student outcomes, along with substantial variation in teachers' effectiveness within schools and districts, makes teacher quality an ideal malleable factor for improving schooling outcomes.

In education, personnel management policies have been implemented for two purposes: to incentivize teachers to refine their instructional practices with the goal of improving student outcomes; and to identify, remediate, and if, necessary, remove the lowest-performing teachers with the goal of improving the quality of the teacher labor force. However, a widely cited report by The New Teacher Project underscored the collective failure of education leaders nationwide to differentiate teacher performance and remediate and remove low-performing teachers (Weisberg et al., 2009). In response, in 2009, the U.S. Department of Education's Race to the

Top (RTTT) competition awarded additional federal aid to states and districts that reformed their approaches to evaluating teachers, setting off a national teacher evaluation reform movement to more accurately identify and dismiss low-performing teachers. In the years that followed, nearly all states (46 of 50 states) and largest school districts (22 of 25 districts) and the District of Columbia (DC) implemented teacher evaluation reforms (Steinberg & Donaldson, 2016). Among the major changes made to evaluation systems were the incorporation of student achievement scores into teacher performance ratings and high-stakes consequences for low-rated teachers, such as remediation plans, tenure revocation, and even termination (Steinberg & Donaldson, 2016).

Recent evidence reveals the potential of teacher evaluation reforms to remove low-performing teachers from the classroom. Evidence from DC and Houston, two districts that revised their teacher evaluation system in the wake of the federal RTTT competition, finds that evaluation reforms can play a meaningful role in the removal of low-performing teachers (Dee & Wyckoff, 2015; Cullen, Koedel, & Parsons, 2016). Additional evidence from DC finds that the increased exit of low-performing teachers has persisted for many years after the initial implementation of the district's evaluation reform (Dee, James, & Wyckoff, 2019). And even in the absence of an explicit policy focused on the removal of low-performing teachers, a pilot evaluation system in Chicago which provided new information to school administrators about their teachers' instructional performance increased the exit of low-rated and non-tenured teachers from the district; however, since explicit accountability sanctions for low-performance were absent from the evaluation pilot in Chicago, there was no commensurate increase in the exit of low-rated tenured teachers (Sartain & Steinberg, 2016). Teacher evaluation reform can also shape the teacher labor market by influencing who enters the teaching force. Evidence suggests

that the overall quality of novice teachers has improved nationally in the wake of evaluation reform, despite the fact that the supply of new teaching candidates has declined over time (Kraft, Brunner, Dougherty, & Schwegman, 2020).

While teacher evaluation reforms have emphasized greater differentiation of teacher performance and greater accountability for low-rated teachers, contracts negotiated between school districts and teachers' unions offer job protections for tenured teachers that may limit the impact that evaluation reforms have on improving the quality of the teacher labor force. For example, in Chicago Public Schools, the focus of this paper, tenured teachers with low evaluation ratings are granted institutional supports and additional time during which their performance is re-evaluated; in contrast, non-tenured teachers can have their contracts non-renewed and therefore be exited from the district at any point during their pre-tenure years. Yet, the organizational logistics of documenting a teacher's low-performance, removing low-performing teachers, and the uncertainty about the quality of the teacher's replacement can be so burdensome that school principals may choose not to pursue this option. In fact, Kraft & Gilmour (2017) find that principals may avoid giving teachers low ratings because of the intensive amount of time required to document the low performance and to implement the professional development and improvement plans that low evaluation ratings typically trigger, especially for tenured teachers. Principals further report that they may avoid dismissing low-performing teachers due to concerns about hiring an even lower-quality replacement teacher from the district's excess pool of tenured teachers (Kraft & Gilmour, 2017).

Recently, teacher tenure protections have been challenged in the courts in states like California, Minnesota, and New York. (See Kraft et al. (2020) for a review of the legal challenges to teacher tenure). In these cases, plaintiffs often argue that the inability to remove

low-performing tenured teachers is unduly onerous, leaving ineffective teachers in the classroom with detrimental effects on student learning. Plaintiffs also cite equity concerns related to teacher tenure protections since disadvantaged students are more likely to be taught by a low-performing teacher; indeed, a host of prior evidence finds that lower-performing teachers tend to be systematically assigned to lower-achieving and higher-poverty schools and students (Allensworth et al., 2009; Clotfelter, Ladd & Vigdor, 2006; Goldhaber, Lavery & Theobald, 2015; Ingersoll, 2001; Kalogrides & Loeb, 2013; Kalogrides, Loeb, & Beteille, 2013; Monk, 1987). In some settings, policy changes have made the path to tenure more difficult or have removed tenure protections altogether. Evidence from New York City and Louisiana indicate that these types of reforms to teacher tenure rules decrease the share of teachers who receive tenure (Loeb, Miller & Wyckoff, 2015) while increasing the exit of less-effective teachers (Loeb et al., 2015; Strunk, Barrett & Lincove, 2017).

In this paper, we examine the interaction of teacher evaluation and the job protections associated with a teacher's tenure status. We do so in the context of Chicago Public Schools (CPS), the nation's third-largest school district, and its teacher evaluation system – Recognizing Educators Advancing Chicago Students (REACH) – which was implemented for the first time in the 2012-13 school year and was still in place as of the 2020-21 school year. CPS is an important context to study the effects of teacher evaluation reform on the exit of low-performing teachers. Chicago's REACH system incorporates high stakes accountability sanctions, including remediation and dismissal, that are tied to the receipt of low evaluation ratings. These design features were incorporated into REACH to better differentiate teacher performance and to increase the accountability function of teacher evaluation for low-performing teachers. And since contractual protections are extended to tenured teachers who receive low evaluation ratings that

are unavailable to low-rated non-tenured teachers (those in their first three years in Chicago), we are able to examine the interaction between teacher evaluation reform and the associated employment protection granted to more experienced teachers in Chicago.

We start by establishing the association between teacher tenure, teacher performance, and teacher exit from Chicago. Non-tenured teachers in Chicago are approximately twice as likely, on average, to exit CPS than their tenured colleagues. Even among teachers whose performance ratings are equivalent in a given year – those who receive the same REACH evaluation rating and the same underlying REACH evaluation score which determines the REACH rating – tenured teachers are significantly less likely than non-tenured teachers to exit CPS, either for any reason or involuntarily. The magnitude of differential exit by tenure status is increasing with lower REACH evaluation ratings, suggesting that tenure provides meaningful employment protections even for the lowest-performing tenured teachers.

Of course, the differential exit by tenure status may reflect variation in the extent to which non-tenured teachers, who have no more than three years of teaching experience in Chicago, are committed to a career in teaching. To avoid conflating a teacher's commitment toward and preferences for a career in teaching with their annual evaluation ratings, we employ a regression discontinuity (RD) design to estimate the effect of evaluation on the exit of low-performing teachers. RD estimates indicate that receipt of an Unsatisfactory rating increased the likelihood that low-rated tenured teachers exit Chicago by the end of the subsequent school year by 50 percent; this substantive increase is driven by the involuntary exit of low-rated tenured teachers from Chicago. And though low-rated non-tenured teachers exit the district at high rates, we do not find that their exit is driven by receipt of low evaluation ratings. We further find no effect of evaluation on the exit of low-rated tenured teachers in the year in which they were evaluated,

suggesting that the timing of the provision of teacher evaluation ratings is consequential for the timing of the exit of low-rated tenured teachers. This is in light of the fact that CPS teachers (and their school administrators) do not receive their formal evaluation ratings from the district until the fall of the subsequent school year (approximately October/November). Thus, the exit of low-rated tenured teachers hinges on the receipt of their formal Unsatisfactory rating; indeed, the consequence of this delay in the receipt of evaluation ratings is that low-performing teachers remained in the classroom for at least an additional school year.

Then, we compare the performance of the low-rated teachers who exited CPS to the performance of the teachers who replaced them in the same school settings. Indeed, if evaluation successfully removes the lowest-performing teachers, it's critical to understand the relative quality of the teacher labor supply available to replace them. We find that replacing low-rated teachers would result in substantive and significant improvements in teacher quality, measured by both classroom observations of a teacher's instructional performance and value-added measures based on student achievement scores, in the schools in which low-rated teachers exited. Given that replacement teachers are considerably higher performing than the Unsatisfactory-rated teachers who exited the same school, policy simulations indicate that the quality of the available teacher labor supply is sufficient to support the removal of many more low-performing teachers in Chicago. We estimate that if the REACH system raised the threshold for an Unsatisfactory rating, this would increase the share of low-rated teachers by nearly fivefold while still realizing improvements in the overall quality of the teacher workforce in Chicago.

Taken together, findings from this paper pose several policy implications. First, this research suggests that teacher evaluation has the potential to improve the quality of the teacher workforce through the identification and removal of low-rated teachers, and low-rated tenured

teachers in particular. Second, our findings suggest that absent high-stakes consequences embedded in Chicago's new evaluation system, low-performing tenured teachers would otherwise remain in the classroom, even though student performance would benefit from those teachers being replaced by teachers who are new to the district. Further, the main limitation of the evaluation reform may be that too few teachers are identified as low performing, given that the available supply of teachers is of considerably higher quality. In the following section, we describe in greater detail the teacher evaluation reform in Chicago Public Schools. We then describe our data and methodology, followed by results, in which we document the distributional consequences of replacing low-rated teachers. We conclude by discussing the implications of these findings and the role of teacher evaluation reform in improving teacher quality via personnel management.

Teacher Evaluation in Chicago Public Schools

We study the effects of teacher evaluation on the exit of low-performing teachers. We do so in the context of REACH, a districtwide teacher evaluation policy implemented in CPS beginning in the 2012-13 school year. The development and implementation of REACH was in response to state legislation in Illinois that required teacher evaluations consist of multiple measures of teacher practice, including classroom observations based on a rubric and indicators of student growth.[1] Prior to REACH, teachers in CPS were evaluated based on a "checklist," where teachers often reported receiving little formal feedback on their performance and in some cases no formal evaluation via classroom observation of their instructional performance (Sartain et al., 2011). Chicago's prior evaluation system, which had been in place since the 1970s, did little to

[1] The Illinois Performance Evaluation Reform Act (PERA) was enacted in 2010 in part to strengthen the state's Race to the Top application. CPS was an early adopter of PERA-based evaluation reform relative to other districts in the state. Prior to PERA, CPS had piloted an informal evaluation system using the Danielson Framework for Teaching to guide classroom observations and pre- and post-observation coaching conversations (Steinberg & Sartain, 2015). This pilot experience helped to inform some of the state legislation, particularly around the use of classroom observations.

differentiate teacher performance, and nearly all teachers received high evaluation ratings. Further, it did not include any measures of student learning. Thus, REACH represented a significant change in how teachers were evaluated and ultimately held accountable for their performance, with the ultimate goal of improving teacher performance and student outcomes.

REACH introduced a mechanism for delivering information about the quality of a teacher's instructional performance to teachers and principals, while codifying new incentives to strengthen the system's accountability function. Certified observers (i.e., the school's principal and/or assistant principal) conduct four formal classroom observations of teachers in their classrooms during the evaluation cycle. Formal observations are followed by a post-observation conference in which the observer provides timely and actionable feedback to teachers. Information about teacher performance is also provided through measures of student growth, though this information is less formative in nature than information provided to teachers during classroom observations of teacher practice. After each evaluation cycle, teachers receive a ratings report that contains their final REACH score and the associated REACH evaluation rating (See Table A1 for more detail on the performance measures and associated weights).[2] The final REACH score is binned into four rating categories that comprise a teacher's formal REACH evaluation rating: Unsatisfactory (100-209 REACH score points), Developing (210-284 REACH score points), Proficient (285-339 REACH score points), and Excellent (340-400 REACH score points). The evaluation cycle differs in length based on a teacher's tenure status and prior evaluation. For high-rated tenured teachers (with prior

[2] There are a few important things to note about how the REACH evaluation score is constructed. First, despite the amount of attention value-added measures receive, approximately 1 in 5 teachers in our sample teach in a tested subject/grade level, so VAMs influence the ratings of relatively few teachers. Second, even for the teachers who do receive VAMs, the VAM itself only accounts for 20 percent of the final score. Third, no high school teachers in the district receive a VAM. And finally, all teachers are rated based on their ability to improve student learning on district-created, subject-specific assessments. Teachers administer and grade their own students' assessments at the beginning and end of the year, and most teachers receive perfect marks on this measure. For all teachers, classroom observation ratings comprise 70 percent of the evaluation rating.

ratings of Proficient or Distinguished), which is the vast majority of tenured teachers, the four required classroom observations occur over a two-year period. For low-rated tenured teachers (with prior ratings of Unsatisfactory or Developing) and all non-tenured teachers, the four observations occur in a single academic year.

REACH also introduced high-stakes consequences tied to the receipt of low ratings (i.e., Unsatisfactory and Developing), and these consequences vary by a teacher's tenure status. Tenured teachers rated Developing receive professional development plans and are subject to annual evaluations until their rating improves to Proficient. Unsatisfactory-rated tenured teachers immediately go under a remediation plan within 30 days of receiving the rating. This remediation plan consists of district and school supports to help teachers improve their practice. These tenured teachers also spend 3-4 hours weekly working with a consultant teacher. After 90 school days, the Unsatisfactory-rated tenured teachers receive another formal REACH evaluation rating based only on classroom observations. If the rating improves to Proficient, the teachers are not subject to layoff; if the rating does not reach proficiency, the teacher may be dismissed. Tenured teachers with "Developing Emerging" ratings (REACH scores in the Developing range and between 210 and 250) for two consecutive years receive an Unsatisfactory rating and are subject to the stakes outlined previously. In contrast, non-tenured teachers with low ratings receive no formal support, and they do not make progress toward attaining tenure status. And further, non-tenured teachers, regardless of performance, are always subject to layoffs and can have their contracts "non-renewed" at will. Non-tenured teachers who receive low ratings do not receive formal support through REACH nor do they make progress toward tenure. There are no rewards associated with receiving high evaluation ratings.

Another important aspect of the REACH evaluation process is the timing of when teacher evaluation data is collected and when teachers and principals receive their formal evaluation ratings (See Table A2 for an overview of the timing of the evaluation cycle). Over the period of data collection, teachers receive feedback and ratings before they are given their final evaluation ratings. In this paper, we refer to that ongoing feedback as informal information about a teacher's instructional practice because teachers do not receive their formal REACH evaluation ratings during this time, but they likely have sufficient information to estimate what their rating will be since classroom observations are heavily weighted in the construction of the final evaluation rating. In fact, it is not until the following fall (usually in October or November) that teachers receive their evaluation ratings and any remediation period begins for low-rated tenured teachers. It is notable that teachers and school administrators do not receive the official evaluation rating until the subsequent year rather than at the end of the evaluation year or even over the summer, as occurs in other contexts. This lag in the receipt of evaluation ratings could have implications for student learning to the extent the low-performing teachers remain in the classroom for an additional year because of the delay in the receipt of official evaluation ratings. In this paper, we explore whether low-performing teachers respond to informal information by exiting the school district prior to receipt of the official evaluation rating.

Data and Sample

We employ administrative data for all CPS teachers in non-charter schools from the 2007-08 through 2018-19 school years. Personnel data include administrative records for individual teachers in each school year and contain information on teacher demographics (race, gender, and birth year), highest level of education attained, National Board certification, and a teacher's tenure

status.[3] These records also include the school where the teacher is employed, allowing us to track movement within and out of the district. Importantly, we also have access to information about the reason for a teacher's exit from CPS. Specifically, these data indicate if a teacher's exit was due to retirement, voluntary resignation, or for an "other reason." We consider teacher exits coded as "other reason" as involuntary – the result of reduction-in-force layoffs, evaluation- and performance-related layoffs, or non-renewal of non-tenured teachers. In our analysis, we label these exits from CPS as "involuntary."

We examine two margins of teacher exit from CPS – any exit and involuntary exit.[4] And, given the timing of the provision of a teacher's formal evaluation rating – teachers evaluated in school year *t* do not receive their formal REACH evaluation rating until fall of school year *t+1* – we consider two time points in which teachers might exit CPS following the evaluation year: exit in year *t* and year *t+1*. Thus, the four outcomes of interest are:

- Any Exit (Year *t*). Any exit at the end of school year *t* would occur after all evaluative classroom observations have been conducted but before teachers have received their formal REACH evaluation rating in the fall of year *t+1*. We interpret this type of exit from CPS as a response to informal information about teacher performance.
- Involuntary Exit (Year *t*). Involuntary exit at the end of school year *t* would also occur after all evaluative classroom observations have been conducted but before teachers have received their formal REACH evaluation rating in the fall of year *t+1*. A teacher's involuntary exit at the end of school year *t* could occur if the information signal received via the classroom observation

[3] CPS teachers earn tenure after being employed for three consecutive years each with an evaluation rating above the Unsatisfactory level. In a National Council on Teacher Quality review of state tenure policies, most states award tenure after three years in the profession (Nitler & Gerber, 2020).

[4] We can also observe within-district transfers in the administrative data. However, we focus on exit from the district because we want to understand the potential of evaluation reform to improve the quality of teaching across the system. Transferring low-rated teachers from one CPS school to another would not result in a shift in the quality distribution of the teacher workforce in Chicago.

process induces school leaders to dismiss low-performing teachers who do not have tenure protections. Yet, we would not expect a tenured teacher to exit CPS involuntarily at the end of year *t* since they have not yet received their formal evaluation ratings.

- Any Exit (Year *t+1*). Any exit at the end of school year *t+1* would occur after teachers received their formal REACH evaluation rating in the fall of year *t+1*. We interpret this type of exit from CPS as a response to the receipt of the formal REACH evaluation rating. For tenured teachers, this type of exit would occur after the completion of a remediation plan associated with the receipt of an Unsatisfactory evaluation rating.
- Involuntary Exit (Year *t+1*). Involuntary exit at the end of school year *t+1* would also occur after teachers received their formal REACH evaluation rating in the fall of year *t+1*. We expect this type of exit from CPS to account for most of the exit of tenured teachers, since tenured teachers who received an Unsatisfactory evaluation rating in year *t* would have completed a remediation plan prior to the end of school year *t+1* that requires tenured teachers to earn a Proficient rating to avoid dismissal from CPS. We later show that, in the context of a regression discontinuity design, tenured teachers rated Unsatisfactory in year *t* perform worse in year *t+1* than tenured teachers rated Developing and whose final REACH score placed them just above the 210 Unsatisfactory/Developing threshold.

The administrative data also include teacher evaluation records from the 2012-13 (the first year of REACH) through 2016-17 school years. Teachers evaluated during the 2016-17 school year would have received their official ratings in fall 2017. The evaluation data include a teacher's formal REACH evaluation rating (Unsatisfactory, Developing, Proficient, or Excellent) from school year *t* and the underlying REACH score (ranging from 100-400) that determines the evaluation rating category. These data also include scores for each of the three components of the

REACH evaluation system – classroom observation scores; value-added measures (VAM) in reading and math; and scores on a district-developed assessment called "performance tasks" (see Table A1). Classroom observation scores, which are on a 1-4 continuous scale, are aggregated across multiple components of teaching and multiple observations that occur during the evaluation cycle. Value-added measures are calculated annually based on student test scores on the NWEA achievement test, are measured in standard deviation units, and are available for teachers of reading and/or math in grades 3-8. Performance tasks were developed to satisfy the state requirement that all teachers have a student growth component to their evaluation. All teachers administer and grade these assessments at the beginning and end of the school year to determine student growth in the subject. Prior research has shown that there is little variation in teachers' performance task scores, with almost all teachers scoring highly on this measure (Jiang & Sporte, 2014).

Sample

Our analytic sample includes all teachers in CPS with a formal REACH evaluation rating in any school year during the 2012-13 through 2016-17 period.[5] While non-tenured teachers are rated annually, high-rated tenured teachers – those who do not receive an Unsatisfactory or Developing rating – receive a formal REACH evaluation rating every other school year. We restrict the sample to teachers who are formally evaluated in any given year and for whom we observe tenure status. Our analytic sample contains 44,637 teacher-by-year observations, of which there are 22,172 unique CPS teachers.

Table 1 summarizes the characteristics of all CPS teachers and the analytic sample, which we also disaggregate by tenure status. Overall in the analytic sample, 76 percent of teachers are female (compared to 77 percent of all CPS teachers); 53 percent of teachers are white (compared

[5] There were 955 teacher-by-year observations where the underlying REACH score was missing but who received a Proficient REACH evaluation rating; we excluded these teachers from the analytic sample.

to 50 percent of all CPS teachers), 21 percent are Black (22 percent of all CPS teachers) and 19 are Latino (20 percent of all CPS teachers). A majority of teachers in the analytic sample – 62 percent – have a graduate degree (compared to 68 percent of all CPS teachers), and 6 percent hold National Board certification (compared to 8 percent of all CPS teachers). Further, 56 percent of the teacher-by-year observations in the analytic sample are of tenured teachers, though 74 percent of all CPS teachers are non-tenured; this is because non-tenured teachers are evaluated annually, while tenured teachers who receive a Proficient or Excellent REACH rating are evaluated every other school year. Compared to tenured teachers in the analytic sample, non-tenured teachers are more likely to be white, and less likely to have a graduate degree or hold National Board Certification. On average, 10 percent of teachers in our analytic sample annually exit CPS (compared to 12 percent of all CPS teachers). Among tenured teachers in our sample, 7 percent annually exit CPS (and 1 percent annually exit involuntarily), while 15 percent of non-tenured teachers annual exit CPS (and 7 percent annually exit involuntarily).

<Table 1 about here>

Table 2 (Panel A) shows the distribution of evaluation ratings for the analytic sample, including the proportion receiving each of the four formal REACH evaluation ratings. Overall, 1 percent of the evaluation ratings received were Unsatisfactory and 20 percent were Developing; non-tenured teachers are more likely to receive Developing ratings (28 percent) than tenured teachers (15 percent). The fact that very few teachers in CPS received Unsatisfactory ratings is consistent with the distribution of teacher evaluation ratings across the country, such as in Michigan, where 0.5 percent of all teachers statewide were rated ineffective, the lowest of four ratings categories, under the state's recently reformed evaluation system (Drake et al., 2019). Panel A also reports the REACH score which underlies the final REACH ratings; teachers' mean

(standard deviation) REACH score was 312.7 (38.8); tenured teachers' mean REACH score was 319.7 (37.8) compared to 303.8 (38.3) for non-tenured teachers. Recall that the ratings threshold below which teachers receive an Unsatisfactory final rating is 210 REACH score points, which is approximated 2.5 standard deviations below the average REACH score among teachers in our sample. In Panel B of Table 2, we report mean performance scores for teachers' classroom observation and VAM (math and reading) scores. For each of the REACH performance measures, tenured teachers received higher scores, on average, than non-tenured teachers.

<Table 2 about here>

Evaluation Ratings, Tenure Status and Teacher Exit

We begin by describing the relationship between teachers' formal REACH evaluation ratings, tenure status, and exit from CPS. Figure 1 shows the likelihood of teacher exit, by tenure status, across the distribution of teacher performance (as measured by the REACH score). Each panel of Figure 1 shows one of the four exit outcomes separately by tenure status, and the vertical lines indicate the three final REACH evaluation rating thresholds: Unsatisfactory/Developing at 210 REACH score points; Developing/Proficient at 285 REACH score points; and Proficient/Excellent at 340 REACH score points. Across the four exit outcomes, non-tenured teachers are always more likely to exit CPS than their tenured colleagues. Specifically, 15 percent of non-tenured teachers annually exit CPS at the end of year *t*, of which 7 percent exited CPS involuntarily; these exit rates for non-tenured teachers compare to an annual exit rate of 7 percent for tenured teachers, of which 1 percent of tenured teachers exited CPS involuntarily. Another key difference in the exit patterns of tenured and non-tenured teachers is the relationship between a teacher's likelihood of exit and their REACH evaluation score. Notably, the REACH score-exit gradient is steeper among non-tenured teachers with REACH evaluation ratings of Unsatisfactory

and Developing than among tenured teachers with the same evaluation rating, suggesting that, among lower-rated teachers, non-tenured teachers are more likely to exit CPS than tenured teachers with the same REACH rating. At the same time, we find no discontinuous change in the likelihood of any exit or involuntary exit at the end of school year *t* at the three evaluation ratings thresholds for tenured and non-tenured teachers (Figure 1, Panels A-D). This is unsurprising given that, by the end of year *t*, teachers have received information about their instructional performance from classroom observations but have yet to receive their final REACH evaluation ratings.

Yet, at the Unsatisfactory/Developing threshold, there is a discontinuous increase in any and involuntary exits at the end of year *t+1* among Unsatisfactory-rated tenured teachers; we do not observe a similar discontinuous jump in teacher exit among Unsatisfactory-rated non-tenured teachers in year *t+1*. This provides descriptive evidence that the labor market outcomes of low-performing teachers depend on both their evaluation ratings and their tenure status.

<Figure 1 about here>

We further explore the relationship between teacher ratings, tenure and exit by estimating variants of the following regression specification:

$$(1)\ Exit_{irst} = \alpha + \beta_1 Tenure_{it} + \sum_{r=1}^{3} \gamma_r Rating_{irt} + \theta_r\left(Tenure_{it} * \sum_{r=1}^{3} Rating_{irt}\right) + f(Score_{it}) + X'_{it}\Gamma + \phi_{st} + \varepsilon_{irst}$$

where $Exit$ equals 1 if teacher *i* with evaluation rating *r* in school *s* exits CPS at the end of school year *t* (or, separately, *t+1*) and zero if teacher *i* remains employed in CPS. In separate regressions, $Exit$ refers to two distinct outcomes – any exit and involuntary exit from CPS. We model exit as a function of a teacher's tenure status (*Tenure*) and a series of indicator variables for a teacher's formal REACH evaluation rating (*Rating*) associated with (though received after the end of) school year *t*; the omitted reference category is the highest REACH rating (i.e., Excellent).

We interact the tenure variable with the vector of indicator variables for REACH evaluation ratings, allowing us to test if there is differential exit between tenured and non-tenured teachers who receive the same evaluation rating. We further control for a flexible function of *Score*, a teacher's underlying REACH score, including linear and quadratic polynomials. *X* is a vector of observable teacher characteristics, including race, gender, birth year, education level and National Board certification. ϕ_{st} is a school-by-year fixed effect that controls for all common shocks experienced by teachers in the same school and in the same academic year, thereby restricting comparisons to teachers teaching in the same school-by-year cell; and ε_{irst} is a random error term.

Table 3 summarizes these results; each column presents one of four exit outcomes. At the end of school year *t* – the year in which a teacher is evaluated but prior to receipt of the formal REACH evaluation rating – teachers rated Unsatisfactory are significantly more likely to exit CPS than higher-rated teachers. On average, teachers rated Unsatisfactory are 14 percentage points more likely to exit CPS for any reason – and 21 percentage points more likely to involuntarily exit CPS – than teachers rated Excellent; teachers rated Developing are, on average, 3.5 percentage points more likely to exit CPS for any reason – and 5.2 percentage points more likely to involuntarily exit CPS – than teachers rated Excellent. There is no difference in exit between teachers rated Proficient and Excellent. Further, the estimated coefficients on the interaction between a teacher's tenure status and evaluation rating provide initial insight into whether tenure status insulates low-rated teachers from exit. We find that tenured teachers are significantly less likely than non-tenured teachers with the same REACH rating and the same REACH score to exit CPS, either for any reason (column 1) or involuntarily (column 2). Notably, the magnitude of differential exit by tenure status is increasing with lower evaluation ratings. Tenured teachers rated Unsatisfactory are 27 percentage points less likely than non-tenured teachers rated Unsatisfactory

to exit CPS, and 41 percentage points less likely to involuntarily exit CPS. By comparison, tenured teachers rated Proficient are 1.6 percentage points less likely than non-tenured teachers rated Proficient to exit CPS, and 1 percentage point less likely to involuntarily exit CPS.

At the end of school year *t+1* – the year in which a teacher receives the formal REACH evaluation rating for school year *t* – teachers rated Unsatisfactory are significantly more likely to exit CPS than higher-rated teachers, and the magnitude of the coefficients associated with any exit in year *t+1* (column 3) are nearly identical to those associated with any exit at the end of school year *t* (column 1). Yet, there is no statistically significant difference in exit between tenured and non-tenured teachers rated Unsatisfactory at the end of year *t+1* – either for any reason (column 3) or involuntarily (column 4) –, suggesting that the evaluation system's formal consequences for low-performance compelled the exit of tenured teachers only after teachers (and their school administrators) received their formal REACH evaluation rating. This is in contrast to the significant difference in exit (any and involuntary) at the end of year *t* between tenured and non-tenured teachers who are rated Unsatisfactory but who have not yet received their formal evaluation rating (columns 1 and 2). We now turn to whether the provision of the formal REACH evaluation rating, and the consequences for low-performance implemented under the REACH evaluation system, increased the exit of low-rated teachers, and the extent to which such exit may have been concentrated among tenured teachers.

<Table 3 about here>

Effects of Evaluation on Teacher Exit

We employ a regression discontinuity (RD) design to estimate the impact of evaluation, and, in particular, the timing of the provision of a teacher's formal evaluation rating, on teacher exit from CPS. To do so, we exploit plausibly exogenous variation in teachers' formal REACH

evaluation ratings induced by discrete differences in the final REACH score around the Unsatisfactory/Developing ratings threshold. While evaluation ratings in Chicago are based on four mutually exclusive ratings categories, a continuous REACH evaluation score underlies the assignment of these ratings. Since teachers rated Unsatisfactory and Developing just below and above the REACH score threshold (i.e., 210 points), respectively, should have, on average, the same observable and unobservable characteristics, we can consider these teachers as good as randomly assigned to formal REACH evaluation ratings (Table A3 presents results testing for discontinuities in teacher characteristics at the various evaluation rating thresholds; we find no consistent evidence of discontinuities in teacher characteristics, particularly at the Unsatisfactory/Developing threshold). We leverage this rating assignment mechanism to estimate the causal effects of REACH ratings (and, importantly, the corresponding dismissal threats) on teacher exit from CPS. Researchers elsewhere have employed a similar strategy to estimate the effect of evaluation reform on teacher turnover, retention, and performance (Dee & Wyckoff, 2015).

We focus on teachers just above/below the Unsatisfactory/Proficient threshold because this is where teachers face high-stakes in terms of remediation and dismissal (see Tables A4 and A5 for results associated with teachers just above/below the Developing/Proficient and Proficient/Excellent thresholds, respectively). We estimate impacts separately for tenured and non-tenured teachers since they are subject to different contractual protections associated with low ratings that might differentially affect teacher exit from CPS. The RD specification takes the following form:

$$(2)\ Exit_{it} = \delta I(REACH_{it} < 0) + f(REACH_{it}) + \gamma(I(REACH_{it} < 0) * f(REACH_{it})) + X'_{it}\Gamma + \lambda_t + \mu_{it}$$

where $Exit$ equals 1 if teacher *i* exits CPS (at all or involuntarily) at the end of school year *t* (or, separately, *t+1*) and zero if teacher *i* remains employed in CPS. *REACH* is the underlying REACH score for teacher *i* in school year *t* that determines a teacher's assignment to a REACH rating, which we center at the relevant threshold (210 points for the contrast between teachers rated Unsatisfactory or Developing; 285 points for the contrast between teachers rated Developing or Proficient; and 340 points for the contrast between teachers rated Proficient or Excellent). We include an indicator function, $I(REACH_{it} < 0)$, that equals 1 if teacher *i* is below the centered REACH score and 0 if teacher *i* is above the centered REACH score, and $f(REACH_{jt})$, which is a smooth function of the a teacher's centered REACH score. We further interact teacher *i*'s REACH score with the indicator function to allow the regression slope to vary on either side of the relevant ratings threshold. The variable λ is a year fixed effect and μ_{it} is a random error term. In alternative specifications of equation (2), we include $\boldsymbol{X}$, which is a vector of teacher characteristics as in equation (1).

We report parametric and nonparametric estimates of the effect of receiving a given final REACH evaluation rating on teacher exit from CPS. For the nonparametric estimates, we use one common mean square error-optimal bandwidth for each outcome separately for tenured and non-tenured teachers using the sharp robust RDD estimator developed by Calonico, Cattaneo, and Titiunik (2014). The coefficient of interest on the indicator function is δ, which captures any shift in teacher exit at the relevant ratings threshold. If, for example, teachers rated Unsatisfactory who are just below the 210 REACH score points threshold are more likely to exit CPS than teachers rated Developing who are just above the 210 REACH score points threshold, then we would expect δ to be positive and significantly different from zero.

Conditions for Causal Inference

The key assumption underlying the internal validity of the RD design is that assignment of teachers to REACH ratings at the ratings threshold is as good as random (Lee & Lemieux, 2010). The extent to which principals or teachers are able to manipulate their REACH score, thus changing teachers' final REACH ratings, poses a threat to this assumption. For example, principals may give struggling teachers the benefit of the doubt and artificially increase their classroom observation scores, which account for the majority of a teacher's final REACH rating. In this way, teachers who should have received an Unsatisfactory rating are moved into the Developing category. If so, this practice would suggest that particular teachers were able to manipulate their formal REACH evaluation rating, calling into question the validity of the RD design. However, evidence from Figure 2, which shows the density of the REACH score by teacher tenure status, indicates that this type of systematic manipulation around the ratings threshold is unlikely to be of concern. Indeed, Figure 2 shows continuity of the assignment variable (i.e., REACH score) at each of the three formal REACH evaluation rating thresholds. Further, we find no evidence of statistically significant discontinuities at any of the evaluation rating thresholds, both for tenured and non-tenured teachers, based on results from a McCrary test (McCrary, 2008). We also provide evidence that the assignment of final REACH evaluation ratings for the analytic sample strictly complied with the rating thresholds outlined in the CPS-CTU contract (see Figure A1). That is, in the analytic sample, all teachers with REACH scores below 210 received an Unsatisfactory evaluation rating.

<Figure 2 about here>

Results

Figure 3 presents graphical evidence on the probability of teacher exit (any and involuntary exit) in years t and $t+1$, by tenure status, as a function of the REACH score at the

Unsatisfactory/Developing threshold (Figures A2 and A3 show the distribution of teacher exit at the Developing/Proficient and Proficient/Excellent thresholds, respectively, by tenure status). At the end of school year *t*, there is no evidence that either tenured teachers (Panels A and C) or non-tenured teachers (Panels B and D) rated Unsatisfactory (and just below the REACH score threshold of 210) exit CPS at higher rates than teachers rated Developing who are just above the 210-point threshold. This is unsurprising because teachers do not receive their formal REACH evaluation ratings until after the start of the next school year. In contrast, by the end of the next school year (i.e., year *t+1*), tenured teachers who are rated Unsatisfactory are much more likely to exit CPS than teachers rated Developing at the 210-point REACH score margin (Figure 3, Panel E). Among Unsatisfactory-rated tenured teachers, the likelihood of any exit in year *t+1* is approximately 0.60; this compares to the likelihood of any exit in year *t+1* of approximately 0.40 for Developing-rated tenured teachers at the margin, representing a 50 percent increase in the exit of Unsatisfactory-rated tenured teachers. This increase in the likelihood of any exit among Unsatisfactory-rated tenured teachers in year *t+1* is very similar in magnitude to the increase in the likelihood of involuntary exit for the same tenured teachers in year *t+1* (Figure 3, Panel G), suggesting that personnel evaluation can induce the exit of low-rated tenured teachers, but only once a teacher receives the binding formal evaluation rating. Thus, even when tenured teachers receive information about their instructional performance during the year of evaluation via ongoing classroom observations by school administrators, tenured teachers are unlikely to exit CPS unless required to do so. For non-tenured teachers, the graphical evidence suggests that the receipt of an Unsatisfactory rating does not induce differential exit from CPS at the end of school year *t+1*. This is consistent with the fact that non-tenured teachers do not have the same employment protections as tenured teachers and can be exited from CPS regardless of their evaluation rating.

<Figure 3 about here>

Next, we generate regression-based estimates of the magnitude and statistical significance of the effect of an Unsatisfactory REACH rating on teacher exit from CPS. We present nonparametric (Table 4) and parametric (Table 5) estimates (at various bandwidths around the evaluation rating threshold) of the effect of receiving an Unsatisfactory rating on teacher exit in years *t* and *t+1*, separately for tenured (Panel A) and non-tenured (Panel B) teachers.[6] In Table 4, nonparametric RD estimates indicate that there is no differential exit from CPS in year *t* – any or involuntary exit – for either tenured or non-tenured teachers. These findings further suggest that low-rated teachers do not respond to informal information about their (low) performance by exiting CPS at the end of year *t*, prior to receiving their formal evaluation ratings, even though teachers know their performance based on classroom observations which largely determine their final REACH rating.

<Table 4 about here>

However, we find consistent and robust evidence that low-rated tenured teachers are much more likely to exit CPS, but only after the receipt of their official REACH evaluation ratings. Nonparametric estimates from Table 4 show a 17.9 percentage point increase in the likelihood of any exit for Unsatisfactory-rated tenured teachers in year *t+1*; this estimate is robust to the inclusion of controls for observable teacher characteristics (Table 4, Panel A, columns 5 and 6), and represents a 50 percent increase in the probability of exit from CPS relative to the counterfactual mean exit rate of 35 percent. Notably, nearly all of the increase in teacher exit is involuntary, as shown in columns (7) and (8) of Table 4. In Table 5, parametric RD results show

[6] While we focus on teachers at the Unsatisfactory/Developing ratings threshold because remediation and dismissal stakes are tied to an Unsatisfactory REACH evaluation rating, we also present nonparametric RD results for teachers at the Developing/Proficient and Proficient/Excellent ratings thresholds (see Tables A4 and A5, respectively).

that the nonparametric RD estimates of teacher exit in year *t+1* are robust across multiple bandwidths of the REACH score for any exit (Table 5, Panel A, columns 5 and 6) and involuntary exit (Table 5, Panel A, columns 7 and 8). For non-tenured teachers without the contract protections of their tenured colleagues, we find no evidence that evaluation increased exit among the lowest-performing teachers by the end of year *t+1*. Taken together, these findings suggest that evaluation reform has played a significant role in relaxing the job protections of low-rated tenured teachers and increasing their exit from Chicago. Indeed, these findings indicate that, in the absence of high-stakes teacher evaluation with binding job dismissal stakes, low-rated tenured teachers would likely remain in the classroom.

<Table 5 about here>

Tenured teachers rated Unsatisfactory are placed on professional development and remediation plans that provide them with additional instructional support. Thus, it is possible that their performance improved and the dismissal of these teachers at the end of the school year would ignore any contemporaneous improvements in performance. To assess whether the performance of Unsatisfactory-rated tenured teachers improved in the subsequent school year, we implement an RD approach similar to above, but in this case the outcome is teacher performance (in year *t+1*) on the final REACH and classroom observation scores, rather than teacher exit. We find that the performance of Unsatisfactory-rated tenured teachers not only didn't improve, but declined in the year after evaluation (i.e., in the year in which they received their formal Unsatisfactory REACH evaluation rating). Compared to tenured teachers just above the 210 REACH score point threshold who received a Developing rating, tenured teachers who received Unsatisfactory evaluation ratings were significantly lower performing (See Table 6). The marginal Unsatisfactory-rated tenured teacher had a REACH score at the end of year *t+1* that was 58 points below the marginal

Developing-rated tenured teacher, which represents approximately a 1.5-standard deviation decline in performance on the REACH score. We further find that the marginal Unsatisfactory-rated tenured teacher's classroom observation scores at the end of year *t+1* were 0.42 points lower at the Unsatisfactory/Developing threshold, corresponding to an approximately 1-standard deviation decline in the measure of instructional performance. Thus, these findings indicate that low-rated tenured teachers were unable to improve their performance even after receipt of their formal evaluation rating and contractually-obligated professional development supports.

<Table 6 about here>

Teacher Labor Supply and the Performance of Replacement Teachers

As we have shown, the REACH evaluation system successfully increased the exit from Chicago of the lowest-performing teachers (though, with a year lag from the year of evaluation). Yet, if the teacher labor supply available to replace exited teachers is no more effective, on average, then the low-rated teachers who exited CPS, then the policy of dismissing low-rated teachers would not have its intended effect – improving the overall distribution of teacher quality in Chicago. In this section, we compare the performance of the low-rated teachers who exited CPS to the performance of those teachers who replaced them. Though we do not observe each exited/replacement pair of teachers (e.g., if a low-rated 5th-grade teacher exited a CPS school, we do not observe the specific teacher who replaced the low-rated teacher in the same 5th-grade classroom in the next school year), we do observe the performance of all teachers who are new to a school in which a low-rated teacher exited at the end of the prior school year.

Table 7 shows the performance scores of Unsatisfactory-rated teachers (in the year in which they received an Unsatisfactory rating) and replacement teachers (in their first year in a school exited by an Unsatisfactory-rated teacher). For replacement teachers, we disaggregate

performance scores for those who are in their first year in a CPS school (*New to CPS*) and those who moved from another school within CPS (*From within CPS*). On average, replacement teachers are much higher performing than Unsatisfactory-rated teachers across multiple teacher performance measures, including the REACH score and the two primary components of the REACH score – classroom observations and VAM. Replacement teachers score 294.9 points, on average, on the REACH score compared to 188.4 REACH score points for Unsatisfactory-rated teachers; this difference corresponds to approximately 2.7 standard deviations of the REACH score. And, while *New to CPS* replacement teachers are slightly lower performing than replacement teachers who moved within CPS to a school with an exited low-rated teacher (*From within CPS*), the performance of *New to CPS* replacement teachers is significantly better than Unsatisfactory-rated teachers, including those who exit at the end of school year *t* and those who remain in CPS in the year after they receive an Unsatisfactory rating (i.e., year *t+1*).

<Table 7 about here>

Different CPS schools likely draw from a different pool of available teachers; thus, we are further interested in the relative performance of the available teacher labor supply across schools from which low-rated teachers exited. To examine this, we compare the performance distribution among replacement teachers who are new to teaching in a CPS school with the performance of low-rated teachers who exited the same CPS school at the end of the prior school year. We formalize this comparison as follows:

$$(3)\ \Delta Quality_{s,t+1}^{AverageReplacement} = \overline{Score}_{js,t+1} - Score_{ist},$$

where $Score_{ist}$ is the performance score (REACH score; classroom observation score; or VAM score) for teacher *i* who received a final REACH rating of Unsatisfactory in school *s* in year

t; and $\overline{Score}_{js,t+1} = \frac{\sum_{j=1}^{J} Score_{js,t+1}}{J}$, where $Score_{js,t+1}$ is the performance score for the j^{th} teacher who is new to school *s* as of year *t+1*. $\Delta Quality_{s,t+1}^{AverageReplacement}$ is the difference in teacher performance between the mean performance of the *j* teachers who entered school *s* in year *t+1* compared to low-rated teacher *i* in school *s* in year *t*. This comparison between replacement and low-rated teachers in the same school assumes that the mean performance represents the quality of the typical teacher that school *s* could recruit in a given school year. We bound this mean difference in the performance distribution by comparing the performance of the lowest-performing new-to-CPS replacement teacher with the performance of the Unsatisfactory-rated teacher (i.e., a lower bound); and an upper bound by comparing the highest-performing new-to-CPS replacement teachers with the Unsatisfactory-rated teacher. We formalize this bounding exercise where equation (4) describes the lower bound of the distribution of the difference in performance and equation (5) describes the upper bound, as follows:

(4) $\Delta Quality_{s,t+1}^{LowestReplacement} = \min\{Score_{js,t+1}\} - Score_{ist}$

(5) $\Delta Quality_{s,t+1}^{HighestReplacement} = \max\{Score_{js,t+1}\} - Score_{ist}$.

For each of the three $\Delta Quality_{s,t+1}$ measures (mean, lower and upper), we aggregate across all Unsatisfactory-rated teachers; these results are shown in Table 8 (and the full distribution is shown in accompanying Figure 4). In Table 8, we present results that include just new-to-CPS replacement teachers in years *t+1* and *t+2*; we show year *t+2* to provide insight into the stability of new-to-CPS teacher quality across school years and relative to the Unsatisfactory-rated teachers in the same school (Table A6 shows results that include all new-to-school replacements – those who move from a school within CPS and new-to-CPS teachers). Independent of how we construct the counterfactual replacement teacher – the lowest-, typical- or highest-rated replacement – the

available labor supply is of much higher quality than Unsatisfactory-rated teachers. This is also true when we restrict the comparison to Unsatisfactory-rated teachers who remain in their schools for at least an extra year because of the delay in receipt of the formal REACH evaluation ratings. Focusing on Panel B of Table 8, we find that Unsatisfactory-rated teachers who remain in the district for an additional year following evaluation (i.e., year $t+1$) score 93 points lower on the REACH score than the typical replacement teacher, approximately 2 standard deviations of the REACH score. Similarly, the typical replacement teacher would represent an improvement in teacher quality of 1.11 and 0.88 teacher-level standard deviations in math and reading VAM, respectively, over the Unsatisfactory-rated teacher who remains in a CPS school for an additional year.

<Table 8 about here>

Figure 4 shows the full distribution of the performance difference between new-to-CPS replacement teachers and Unsatisfactory-rated teachers. In each panel, the vertical line is drawn at 0, indicating that the average replacement teacher in the same school is performing at the same level as the Unsatisfactory-rated teacher. For each performance measure, the mass of the distribution lies to the right of the vertical line, suggesting that there are large teacher quality gains from replacing Unsatisfactory-rated teachers with teachers from the available teacher labor supply.

<Figure 4 about here>

Policy Simulation: Changing the Performance Standard for Unsatisfactory Teaching

Thus far, we have established that replacement teachers are considerably higher performing than Unsatisfactory-rated teachers located in the same school; as a result, the likelihood of replacing an Unsatisfactory-rated teacher with a lower-performing teacher is quite low. Specifically, based on the REACH score, replacing an Unsatisfactory-rated teacher with an

average new-to-CPS teacher in the same school would result in the position being filled by a lower-performing teacher in only one case (of 355 cases). However, the extent to which the evaluation policy can shift the overall distribution of teacher quality in Chicago is limited by the fact that just 1 percent of CPS teachers are identified as Unsatisfactory in any given year. In this section, we consider the implications of changing the threshold for an Unsatisfactory rating in ways that result in a greater share of CPS teachers rated Unsatisfactory. Indeed, if CPS district officials aimed to increase the share of teachers rated Unsatisfactory by raising the REACH score threshold, what would be the implications of this change for the distribution of teacher performance? And further, can the available teacher labor supply support such a policy change?

To address this, we examine the distribution of teacher performance scores below three different REACH score thresholds: 210 points (the current Unsatisfactory rating threshold); 230 points; and 250 points. Table 9 presents the distribution of REACH ratings, as well as the final REACH score and performance measures (classroom observation and VAM scores), at each of these three thresholds (we focus discussion here on tenured teachers; results for non-tenured teachers, which are qualitatively similar, are also presented in Table 9). Under the current REACH system (i.e., 210 points), 239 tenured teachers received Unsatisfactory ratings during the 2012-13 through 2016-17 study period. Yet, at a ratings threshold of 230 points, 417 tenured teachers would have received an Unsatisfactory rating; and, at a threshold of 250 points, 900 tenured teachers would have been rated Unsatisfactory. As expected, teachers' performance scores are monotonically increasing as the REACH score threshold increases, and the share of tenured teachers rated Unsatisfactory would increase from 1 percent (at 210) to 2 percent (at 230) to just 4 percent (at 250). Moreover, even at a threshold of 250 points, the performance scores of tenured teachers who would be rated Unsatisfactory would remain well below the districtwide mean.

Specifically, the mean REACH score of 222.5 points among Unsatisfactory-rated tenured teachers at the 250-point threshold is approximately 2 standard deviations lower than the districtwide mean of 312.7 REACH score points (see Table 2). We similarly find significant differences in the classroom observation and VAM scores of tenured teachers below a 250-point threshold compared to the CPS districtwide mean.

<Table 9 about here>

Lastly, we examine whether (and the extent to which) the available teacher labor supply, as defined by new-to-CPS replacement teachers, is higher performing than the low-performing teachers identified at each of the three Unsatisfactory-rating thresholds. Figure 5 shows the distribution of the difference in performance scores between new-to-CPS replacement and teachers who would be rated Unsatisfactory at each of the three Unsatisfactory-rating thresholds. We again note that we restrict the performance comparison of replacement and Unsatisfactory-rated teachers to within the same school. In Figure 5, a value of 0 indicates that the average replacement teacher has the same performance as the Unsatisfactory-rated teacher in the same school. In Panel A, which shows the distribution of the difference in REACH scores, we find that at the 250-point threshold, 91 percent of distribution is to the right of 0, though the mass of the distribution in this case is closer to 0 than at lower thresholds. This pattern by which the overwhelming share of new-to-CPS teachers is higher-performing than Unsatisfactory-rated teachers holds across the different performance measures, and indicates that the quality of the available teacher labor supply is sufficient to accommodate changing the threshold for an Unsatisfactory performance rating.

<Figure 5 about here>

Conclusion

In this paper, we examined the impact of teacher evaluation reform on the exit of low-performing teachers from Chicago Public Schools, with particular interest on the potentially differential effect by a teacher's tenure status. Indeed, tenured teachers in Chicago benefit from contractual protections unavailable to their non-tenured colleagues. Tenured teachers who receive Unsatisfactory ratings under the REACH evaluation system are provided intensive professional development and support and are afforded an opportunity to demonstrate instructional improvement prior to facing dismissal. In contrast, non-tenured teachers can have their contracts terminated at any time. Though we find that tenured teachers are significantly less likely, on average, to exit Chicago than equivalently rated non-tenured teachers, regression discontinuity estimates indicate that receipt of an Unsatisfactory evaluation rating increased the exit of tenured teachers from Chicago by 50 percent, and this increase is driven by their involuntary exit from the district. And while low-rated non-tenured teachers exit Chicago at high rates, we find no evidence that the evaluation system itself induced that exit.

Notably, our findings reveal that the timing of the provision of a teacher's evaluation rating is consequential for determining when low-rated tenured teachers exit the district. Under Chicago's REACH system, teachers and their school administrators do not receive final teacher evaluation ratings until well into the fall of the subsequent school year. This contrasts with teacher evaluation systems in other urban districts, such as DC Public Schools, where teachers receive their final evaluation ratings prior to the start of the next school year (Dee & Wyckoff, 2015). One consequence of this evaluation system feature is that low-rated teachers, particularly tenured teachers, will remain in the classroom for at least an additional year. At the same time, this feature also affords us the opportunity to examine whether teachers who will receive

Unsatisfactory ratings voluntarily exit the classroom in response to information about their performance received during the annual evaluation process (i.e., from observations of their classroom instruction), or delay exit for a year and exit involuntarily only after receiving their official evaluation rating. We find that the increase in exit of low-rated tenured teachers occurs only after the receipt of the official rating, and is driven almost entirely by their involuntary exit (i.e., dismissal) from the district.

Yet, critics of high-stakes teacher evaluation systems may argue that low-performing teachers should be provided with an opportunity to improve their instructional performance. Due to the time lag in receipt of a teacher's final evaluation rating in Chicago, we are able to examine this concern. We find that the instructional performance of Unsatisfactory-rated tenured teachers who remain in Chicago for an additional year not only doesn't improve, but declines compared to teachers just above the district-determined threshold for an Unsatisfactory rating. Moreover, school administrators may avoid dismissing low-performing teachers due to concerns about whether the available teacher supply is of sufficient quality to replace exited teachers (Kraft & Gilmour, 2017). Evidence herein indicates this concern is unfounded. We find that the instructional quality of the available teacher labor supply in Chicago is sufficient not only to support replacing existing low-rated teachers, but also to expand the share of teachers in Chicago receiving Unsatisfactory ratings and therefore subject to dismissal.

Finally, as reforms to teacher evaluation systems have rolled out across the country, there are still ongoing concerns that too few teachers are annually identified for instructional improvement or removal from the classroom for low performance. A systematic review of states that have recently implemented teacher evaluation reforms finds that less than one percent of teachers have been identified as low-performing (Kraft & Gilmour, 2017), indicating that the

identification of low-performing teachers has changed little in the decade since the national movement to reform teacher evaluation began. In fact, one of the barriers to improving the quality of the teacher workforce via personnel management is the continued lack of identification of low-performing teachers whose practices are detrimental to student learning.. This fact is consistent in Chicago, where we show that fewer than 1 percent of teachers annually are identified as Unsatisfactory. Evidence from this paper shows that while the potential for evaluation systems to shift the distribution of teacher quality has yet to be fully realized, changes to existing evaluation policies can accomplish this by changing the performance standard for unsatisfactory teaching.

Ultimately, our findings reveal the important role that the design of evaluation systems play in determining both who is deemed low-performing and when low-performing teachers are subject to dismissal. Thus, education leaders and policymakers in districts like Chicago and elsewhere should consider refining two important design features of teacher evaluation systems – the standard for low-performance and the timing of evaluation ratings. In doing so, systems of evaluation may successfully satisfy their two primary objectives – improving teacher quality and student achievement.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, *25*(1), 95-135.

Allensworth, E., Ponisciak, S., & Mazzeo, C. (2009). *The schools teachers leave: Teacher mobility in Chicago Public Schools*. The University of Chicago Consortium on Chicago School Research.

Calonico, S., Cattaneo, M.D., Farrell, M.H., & Titiunik, R. (2017). Rdrobust: Software for regression-discontinuity designs. *Stata Journal, 17*(2), 372–404.

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

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review*, *104*(9), 2633-79.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2006). Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources*, *41*(4), 778-820.

Cullen, J. B., Koedel, C., & Parsons, E. (2016). The compositional effect of rigorous teacher evaluation on workforce quality. *Education Finance and Policy*, 1-85.

Dee, T. S., James, J., & Wyckoff, J. (2019). Is Effective Teacher Evaluation Sustainable? Evidence from DCPS. *Education Finance and Policy*, 1-53.

Dee, T. S., & Wyckoff, J. (2015). Incentives, selection, and teacher performance: Evidence from IMPACT. *Journal of Policy Analysis and Management*, *34*(2), 267-297.

Drake, S., Auletto, A., & Cowen, J.M. (2019). Grading teachers: Race and gender differences in low evaluation ratings and teacher employment outcomes. *American Educational Research Journal*, 56(5), 1800-1833.

Goldhaber, D. (2002). The mystery of good teaching. *Education Next*, *2*(1), 50-55.

Goldhaber, D., Lavery, L., & Theobald, R. (2015). Uneven playing field? Assessing the teacher quality gap between advantaged and disadvantaged students. *Educational Researcher*, *44*(5), 293-307.

Ingersoll, R. M. (2001). Teacher turnover and teacher shortages: An organizational analysis. *American Educational Research Journal*, *38*(3), 499-534.

Jackson, K., Porter, S., Easton, J., Blanchard, A., & Kiguel, S. (In Press). School effects on socio-emotional development, school-based arrests, and educational attainment. *American Economic Review: Insights*.

Jiang, J.Y., & Sporte, S.E. (2014). *Teacher evaluation in practice: Year 2 teacher and administrator perceptions of REACH.* The University of Chicago Consortium on Chicago School Research

Kalogrides, D., & Loeb, S. (2013). Different teachers, different peers: The magnitude of student sorting within schools. *Educational Researcher*, *42*(6), 304-316.

Kalogrides, D., Loeb, S., & Béteille, T. (2013). Systematic sorting: Teacher characteristics and class assignments. *Sociology of Education*, *86*(2), 103-123.

Kraft, M. A., Brunner, E. J., Dougherty, S. M., & Schwegman, D. J. (2020). Teacher accountability reforms and the supply and quality of new teachers. *Journal of Public Economics*, 188, 104212.

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.

Kraft, M. A. (2019). Teacher effects on complex cognitive skills and social-emotional competencies. *Journal of Human Resources*, *54*(1), 1-36.

Loeb, S., Miller, L. C., & Wyckoff, J. (2015). Performance screens for school improvement: The case of teacher tenure reform in New York City. *Educational Researcher*, *44*(4), 199-212.

Nitler, K., & Gerber, N. (2020, March 12). *Tenure decisions and teacher effectiveness.* National Council on Teacher Quality. https://www.nctq.org/blog/Tenure-decisions-and-teacher-effectiveness

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

Monk, D. H. (1987). Assigning elementary pupils to their teachers. *The Elementary School Journal*, 88(2), 166–187.

Murnane, R. J., & Steele, J. L. (2007). What is the problem? The challenge of providing effective teachers for all children. *The Future of Children*, 15-43.

National Commission on Excellence in Education. (1983). A nation at risk: The imperative for educational reform. *The Elementary School Journal*, *84*(2), 113-130.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417-458.

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.

Sartain, L., Stoelinga, S. R., & Brown, E. R. (2011). *Rethinking teacher evaluation in Chicago: Lessons learned from classroom observations, principal-teacher conferences, and district implementation.* The University of Chicago Consortium on Chicago School Research.

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

Steinberg, 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.

Strunk, K. O., Barrett, N., & Lincove, J. A. (2017). *When tenure ends: The short-run effects of the elimination of Louisiana's teacher employment protections on teacher exit and retirement.* Education Research Alliance for New Orleans.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness.* The New Teacher Project.

Tables & Figures

Table 1. Teacher Characteristics

		Analytic Sample		
	All CPS Teachers	**All Teachers**	**Tenured Teachers**	**Non-tenured Teachers**
Female	0.77	0.76	0.77	0.74
Black	0.22	0.21	0.25	0.16
Latino	0.20	0.19	0.21	0.17
White	0.50	0.53	0.47	0.60
Asian/Other	0.08	0.07	0.07	0.08
Graduate Degree	0.68	0.62	0.73	0.49
National Board Certified	0.08	0.06	0.10	0.01
Tenured	0.74	0.56	1.00	0.00
Birth Year	1973.2 (11.1)	1975.8 (11.0)	1970.6 (10.1)	1982.4 (8.1)
Exit CPS	0.12	0.10	0.07	0.15
Involuntary Exit from CPS	0.04	0.04	0.01	0.07
Teachers	26,730	22,172	14,824	9,905
Teacher*Year Observations	96,491	44,637	24,968	19,669

Notes. Each cell reports proportion, except birth year which reports mean (standard deviation). Data are for the 2012-13 through 2016-17 school years. Data include Chicago Public School teachers present in any year during the study period (charter and alternative school teachers are excluded). *Graduate Degree* includes teachers with a master's or doctorate degree.

Table 2. Teacher Evaluation Ratings, by Tenure Status

	All Teachers	Tenured Teachers	Non-tenured Teachers
Panel A: Final Ratings			
Unsatisfactory	0.01	0.01	0.01
Developing	0.20	0.15	0.28
Proficient	0.53	0.52	0.54
Excellent	0.25	0.32	0.18
REACH Score	312.7 (38.8)	319.7 (37.8)	303.8 (38.3)
Panel B: Performance Measures			
Classroom Observation	3.14 (0.45)	3.22 (0.44)	3.03 (0.44)
VAM (Math)	0.02 (0.89)	0.05 (0.87)	-0.01 (0.92)
VAM (Reading)	0.02 (0.79)	0.06 (0.75)	-0.04 (0.83)
Teachers	22,172	14,824	9,905
Teacher*Year Observations	44,637	24,968	19,669

Notes. Each cell reports mean (standard deviation), except Final Ratings categories, which report proportions. *REACH Score* is the teacher's summative evaluation score (based on multiple performance measures - see Table A1) upon which a teacher's formal REACH evaluation rating is based and is on a 100-400 continuous point scale. Teachers whose *REACH Score* is below 210 receive an Unsatisfactory rating; teachers whose *REACH Score* is 210-284 receive a Developing rating; teachers whose *REACH Score* is 284-339 receive a Proficient rating; and teachers whose *REACH Score* is greater than 339 receive an Excellent rating. A teacher's *Classroom Observation* score is based on multiple classroom observations of a teacher's instruction, and is measured on a 1-4 integer scale. A teacher's VAM score is based on a teacher's contribution to student achievement growth (in math or reading).

Table 3. Association between Teacher Tenure, Evaluation Ratings and Exit from CPS

	Any Exit (Year t)	Involuntary Exit (Year t)	Any Exit (Year t+1)	Involuntary Exit (Year t+1)
	(1)	(2)	(3)	(4)
Tenured	-.030***	.003	-.067***	-.002
	(.006)	(.003)	(.008)	(.004)
Unsatisfactory	.143***	.210***	.134***	.049
	(.040)	(.036)	(.042)	(.038)
Developing	.035***	.052***	.054***	.026***
	(.012)	(.007)	(.015)	(.008)
Proficient	.010	.004	.008	.005
	(.008)	(.004)	(.010)	(.005)
Tenured*Unsatisfactory	-.267***	-.410***	-.071	.043
	(.043)	(.037)	(.044)	(.046)
Tenured*Developing	-.090***	-.118***	-.067***	-.030***
	(.010)	(.006)	(.013)	(.007)
Tenured*Proficient	-.016**	-.010***	-.020**	-.010**
	(.007)	(.003)	(.009)	(.004)
P-value from F-test: *Unsatisfactory= Developing=Proficient*	.000	.000	.136	.025
P-value from F-test: *Tenure*Unsatisfactory= Tenure*Developing= Tenure*Proficient*	.000	.000	.000	.004
Adjusted R^2	0.095	0.181	0.097	0.049
Teacher*Year Observations	44,637	44,637	44,637	28,214

Notes. Coefficients reported with robust standard errors (clustered at the school level). All regressions include school-by-year fixed effects, linear and quadratic polynomials in the REACH Score, and the following teacher characteristics: race, gender, birth year, education level and National Board certification. *Any Exit (Year t)* includes teachers who exited Chicago Public Schools (CPS) at the end of the current school year*; Any Exit (Year t+1)* includes teachers who exited CPS at the end of the subsequent school year. *Involuntary Exit* includes teachers who exited CPS for reasons other than retirement or resignation. The omitted reference category includes teachers who were rated Excellent in a given school year. Coefficients are statistically significant at the *10%, **5% and ***1% levels.

Table 4. Nonparametric RD Estimates of the Impact of Unsatisfactory Evaluation Rating on Teacher Exit, by Tenure Status

	Any Exit (Year t)		Involuntary Exit (Year t)		Any Exit (Year t+1)		Involuntary Exit (Year t+1)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Tenured								
Unsatisfactory (relative to Developing)	-.078 (.089)	-.080 (.094)	-.046 (.072)	-.058 (.072)	.179* (.099)	.179* (.101)	.153** (.070)	.137** (.067)
Counterfactual Mean	0.22	0.22	0.08	0.08	0.35	0.36	0.05	0.05
Bandwidth	28.1	25.3	21.5	20.8	29.9	28.0	25.0	25.0
N (left)	179	164	143	137	187	179	102	102
N (right)	409	339	243	227	451	407	223	223
Panel B: Non-tenured								
Unsatisfactory (relative to Developing)	-.021 (.094)	-.008 (.094)	-.017 (.109)	-.008 (.110)	-.022 (.079)	-.010 (.081)	.041 (.074)	.078 (.095)
Counterfactual Mean	0.43	0.43	0.38	0.38	0.53	0.54	0.10	0.11
Bandwidth	18.6	18.6	14.3	14.0	24.2	22.8	25.0	25.0
N (left)	168	169	138	135	200	194	91	91
N (right)	420	421	299	285	627	585	376	376
Year FE	X	X	X	X	X	X	X	X
Teacher Xs		X		X		X		X

Notes. Each column (within a panel) is a separate regression. Coefficients from nonparametric regression discontinuity (RD) reported with robust standard errors (clustered at the school level). All regressions include controls for the linear running variable – a teacher's final REACH score (from year *t*). *Teacher Xs* include controls for teacher gender, race/ethnicity, educational attainment, National Board Certification, and birth year. Coefficients are statistically significant at the *10%, **5% and ***1% levels.

Table 5. Parametric RD Estimates of the Impact of Unsatisfactory Evaluation Rating on Teacher Exit, by Tenure Status

	Any Exit (Year t)		Involuntary Exit (Year t)		Any Exit (Year t+1)		Involuntary Exit (Year t+1)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Tenured								
Unsatisfactory (relative to Developing) (BW=20)	-.090 (.102) [350]	-.109 (.107) [350]	-.087 (.070) [350]	-.108 (.073) [350]	.140 (.109) [350]	.130 (.110) [350]	.199** (.094) [232]	.185* (.096) [232]
Unsatisfactory (relative to Developing) (BW=25)	-.104 (.084) [495]	-.129 (.083) [495]	-.058 (.063) [495]	-.070 (.063) [495]	.175* (.096) [495]	.154* (.092) [495]	.173* (.080) [325]	.162* (.083) [325]
Unsatisfactory (relative to Developing) (BW=30)	-.097 (.074) [651]	-.117 (.074) [651]	-.052 (.054) [651]	-.060 (.054) [651]	.175** (.084) [651]	.153* (.081) [651]	.185** (.073) [426]	.171** (.075) [426]
Panel B: Non-tenured								
Unsatisfactory (BW=20)	-.079 (.079) [656]	-.075 (.079) [656]	-.121 (.078) [656]	-.114 (.078) [656]	-.060 (.085) [656]	-.059 (.086) [656]	.025 (.085) [344]	.023 (.085) [344]
Unsatisfactory (BW=25)	-.025 (.074) [867]	-.018 (.072) [867]	-.064 (.073) [867]	-.058 (.073) [867]	-.017 (.075) [867]	-.012 (.076) [867]	.035 (.077) [468]	.036 (.077) [468]
Unsatisfactory (BW=30)	-.017 (.068) [1083]	-.006 (.067) [1083]	-.037 (.070) [1083]	-.025 (.070) [1083]	-.009 (.070) [1083]	.001 (.070) [1083]	.071 (.069) [600]	.081 (.069) [600]
Year FE	X	X	X	X	X	X	X	X
Teacher Xs		X		X		X		X

Notes. Each cell (within a column and panel) is a separate regression Coefficients from parametric regression discontinuity (RD) reported with robust standard errors (clustered at the school level) in parentheses and sample size in brackets. All regressions include controls for the linear running variable – a teacher's final REACH score (from year *t*) centered around the 210 threshold for Unsatisfactory/Developing – and the centered running variable interacted with the Unsatisfactory indicator. *Teacher Xs* include controls for teacher gender, race/ethnicity, educational attainment, National Board Certification, and birth year. Coefficients are statistically significant at the *10%, **5% and ***1% levels.

Table 6. Nonparametric RD Estimates of the Impact of Unsatisfactory Evaluation Rating on Subsequent REACH Score, by Tenure Status

	REACH Score (Year t+1)		Classroom Observation Score (Year t+1)	
	(1)	(2)	(3)	(4)
Panel A: Tenured				
Unsatisfactory (relative to Developing)	-61.4**	-57.5***	-0.46*	-0.42**
	(27.4)	(20.5)	(0.24)	(0.21)
Counterfactual Mean	257.0	253.7	2.25	2.60
Bandwidth	7.9	8.9	8.6	9.4
N (left)	37	40	41	42
N (right)	33	38	46	52
Panel B: Non-tenured				
Unsatisfactory (relative to Developing)	-9.2	-11.8	-0.24	-0.13
	(16.5)	(18.3)	(0.19)	(0.21)
Counterfactual Mean	271.6	272.4	2.25	2.59
Bandwidth	14.3	11.7	11.6	11.0
N (left)	44	40	51	47
N (right)	127	99	115	105
Year FE	X	X	X	X
Teacher Xs		X		X

Notes. Each column (within a panel) is a separate regression. Coefficients from nonparametric regression discontinuity (RD) reported with robust standard errors (clustered at the school level). All regressions include controls for the linear running variable – a teacher's final REACH score (from year *t*). *Teacher Xs* include controls for teacher gender, race/ethnicity, educational attainment, National Board Certification, and birth year. Coefficients are statistically significant at the *10%, **5% and ***1% levels.

Table 7. Performance Measures for Replacement Teachers

	Replacement Teachers			Unsatisfactory-Rated Teachers		
	All	**New to CPS**	**From within CPS**	**All**	**Exit (Year t)**	**Remain (Year t+1)**
Panel A: Final Ratings						
Unsatisfactory	0.03	0.02	0.03	1.00	1.00	1.00
Developing	0.35	0.38	0.30	0.00	0.00	0.00
Proficient	0.52	0.51	0.54	0.00	0.00	0.00
Excellent	0.10	0.08	0.13	0.00	0.00	0.00
REACH Score	294.89 (39.23)	290.25 (37.93)	300.8 (40.05)	188.42 (19.30)	185.74 (20.81)	191.00 (17.39)
Panel B: Performance Measures						
Classroom Observation	2.89 (0.48)	2.83 (0.46)	2.96 (0.50)	1.80 (0.28)	1.75 (0.29)	1.84 (0.26)
VAM (Math)	-0.10 (0.98)	-0.13 (0.97)	-0.07 (1.00)	-1.05 (1.02)	-1.15 (0.97)	-0.98 (1.06)
VAM (Reading)	-0.07 (0.92)	-0.14 (0.88)	0.01 (0.94)	-0.92 (1.04)	-0.81 (1.11)	-1.00 (0.98)
Teachers	12,126	7,548	5,797	537	263	274
Teachers*Year Observations	14,463	7,548	6,915	537	263	274

Notes. Each cell reports mean (standard deviation), except Final Ratings categories, which report proportions. *REACH Score* is the teacher's summative evaluation score (based on multiple performance measures - see Table A1) upon which a teacher's formal REACH evaluation rating is based and is on a 100-400 continuous point scale. A teacher's *Classroom Observation* score is based on multiple classroom observations of a teacher's instruction, and is measured on a 1-4 integer scale. A teacher's VAM score is based on a teacher's contribution to student achievement growth (in math or reading).

Table 8. Performance Comparison of Unsatisfactory-Rated Teachers to New-to-CPS Replacement Teachers in the Same School

	New to School (Year t+1)			New to School (Year t+2)		
	Lowest Rated	Average Rated	Highest Rated	Lowest Rated	Average Rated	Highest Rated
Panel A: *Unsatisfactory Teachers*						
REACH Score	65.61 (46.39)	93.99 (37.02)	118.97 (40.13)	69.96 (49.89)	98.91 (38.47)	124.66 (39.22)
Classroom Observation	0.55 (0.60)	0.92 (0.48)	1.23 (0.50)	0.67 (0.61)	0.98 (0.48)	1.25 (0.50)
VAM (Math)	0.80 (1.31)	0.99 (1.29)	1.20 (1.37)	0.86 (1.21)	0.93 (1.22)	1.05 (1.25)
VAM (Reading)	0.52 (1.04)	0.82 (1.04)	1.13 (1.30)	0.64 (1.51)	0.87 (1.34)	1.13 (1.27)
Panel B: *Unsatisfactory Teachers Who Remain in t+1*						
REACH Score	65.75 (45.21)	93.42 (33.78)	116.77 (36.72)	72.65 (45.71)	98.72 (34.96)	123.12 (34.68)
Classroom Observation	0.52 (0.62)	0.87 (0.48)	1.17 (0.49)	0.71 (0.55)	0.97 (0.42)	1.21 (0.44)
VAM (Math)	0.98 (1.35)	1.11 (1.42)	1.29 (1.60)	1.22 (1.31)	1.28 (1.37)	1.36 (1.44)
VAM (Reading)	0.60 (0.85)	0.88 (0.89)	1.16 (1.16)	0.96 (1.41)	1.06 (1.33)	1.16 (1.28)

Notes. Each cell reports the mean (standard deviation) difference in teacher performance (by performance measure) between CPS teachers rated Unsatisfactory (in school year t) and teachers who are new to the same school in subsequent school years (either year t+1 or year t+2). In all cells, positive values indicate that the new-to-CPS replacement teachers are higher performing than the Unsatisfactory-rated teachers.

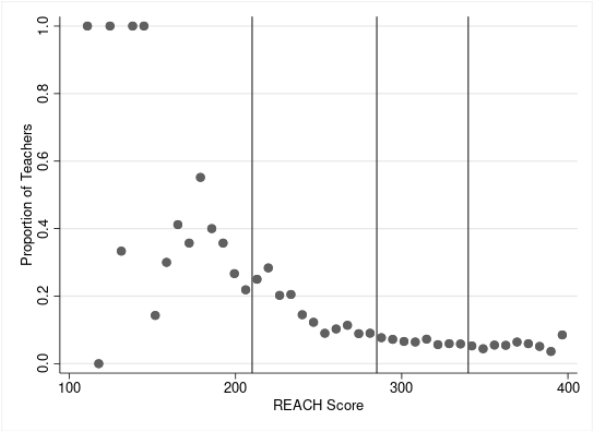
Table 9. Teacher Evaluation Ratings, by Policy Threshold for Unsatisfactory Rating

	Tenured			Non-Tenured		
	REACH Score < 210	**REACH Score < 230**	**REACH Score < 250**	**REACH Score < 210**	**REACH Score < 230**	**REACH Score < 250**
Panel A: Final Ratings						
Share of All Teachers Rated Unsatisfactory	0.01	0.02	0.04	0.01	0.04	0.08
Unsatisfactory	1.00	0.56	0.26	1.00	0.36	0.16
Developing	0.00	0.44	0.74	0.00	0.64	0.84
Proficient	0.00	0.00	0.00	0.00	0.00	0.00
Excellent	0.00	0.00	0.00	0.00	0.00	0.00
REACH Score	184.9 (21.22)	200.6 (24.28)	222.53 (26.26)	192.2 (16.28)	210.8 (17.65)	227.9 (19.45)
Panel B: Performance Measures						
Classroom Observation	1.75 (0.29)	1.91 (0.32)	2.16 (0.36)	1.84 (0.26)	2.03 (0.27)	2.23 (0.30)
VAM (Math)	-0.98 (0.87)	-0.85 (0.93)	-0.74 (0.91)	-1.10 (1.11)	-0.78 (1.01)	-0.67 (0.95)
VAM (Reading)	-0.97 (1.02)	-0.76 (1.02)	-0.55 (1.02)	-0.88 (1.06)	-0.77 (0.92)	-0.63 (0.90)
Teachers	239	417	900	250	688	1,482
Teacher*Year Observations	275	487	1,069	262	730	1,679

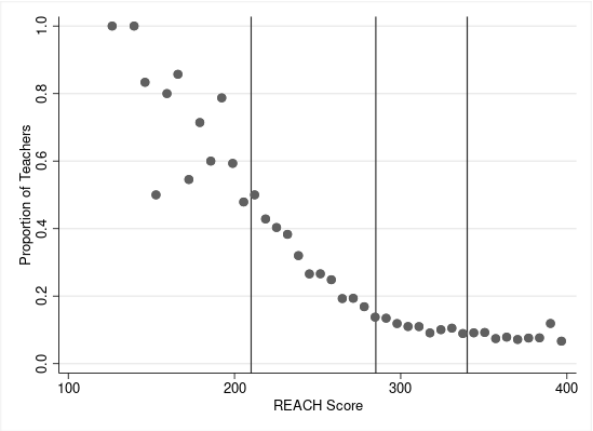
Notes. Each cell reports mean (standard deviation), except Final Ratings categories, which report proportions, for different *REACH Score* thresholds determining teacher assignment to an Unsatisfactory annual REACH evaluation rating.

Figure 1. Likelihood of Teacher Exit from CPS, by Year and Tenure Status

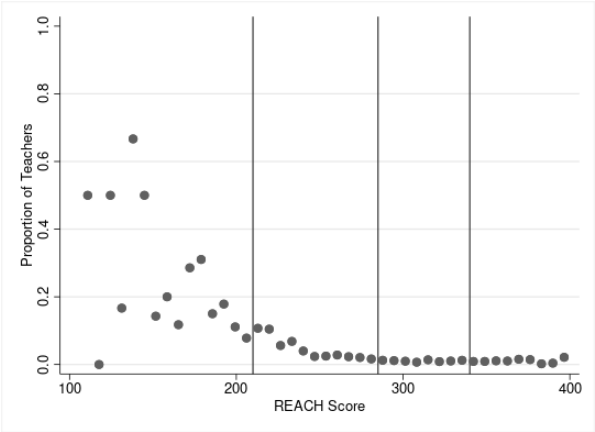
Panel A. Any Exit (Year *t*), Tenured



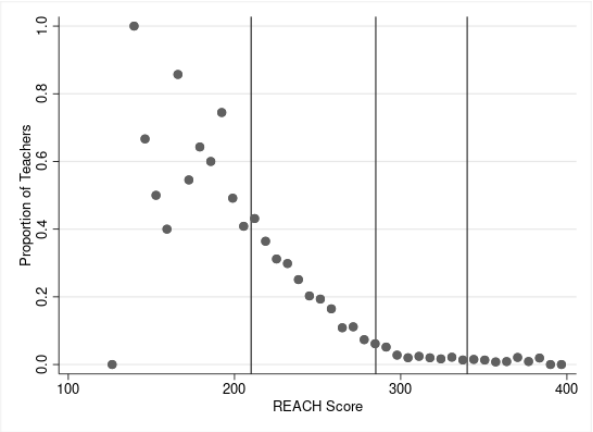
Panel B. Any Exit (Year *t*), Non-Tenured



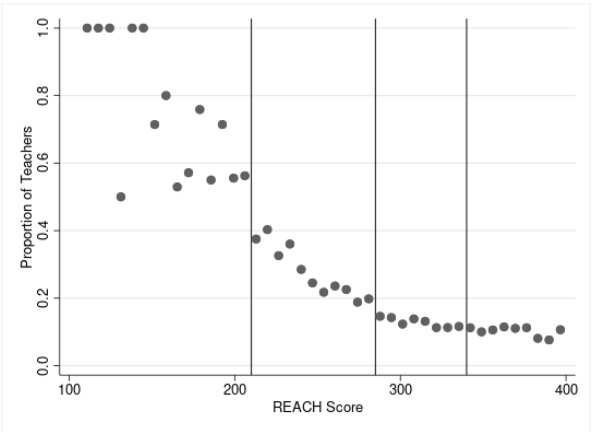
Panel C. Involuntary Exit (Year *t*), Tenured



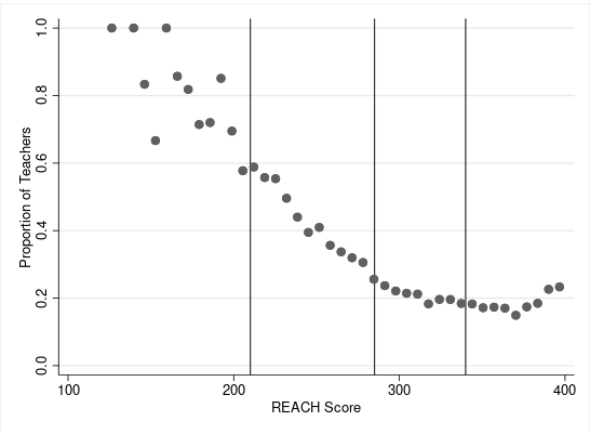
Panel D. Involuntary Exit (Year *t*), Non-Tenured



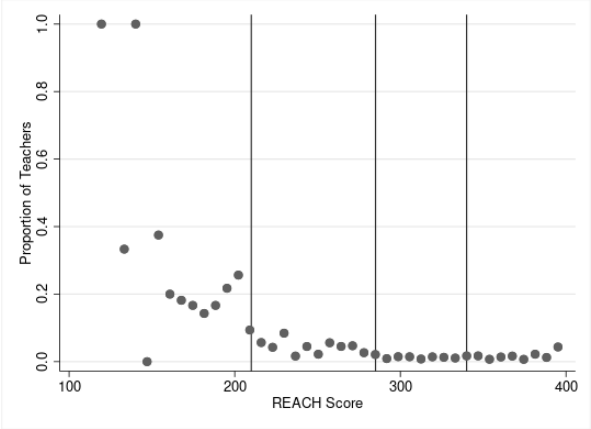
Panel E. Any Exit (Year *t+1*), Tenured



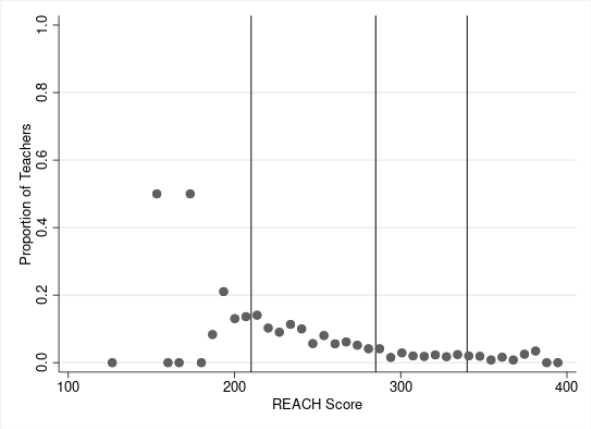
Panel F. Any Exit (Year *t+1*), Non-Tenured



Panel G. Involuntary Exit (Year *t+1*), Tenured

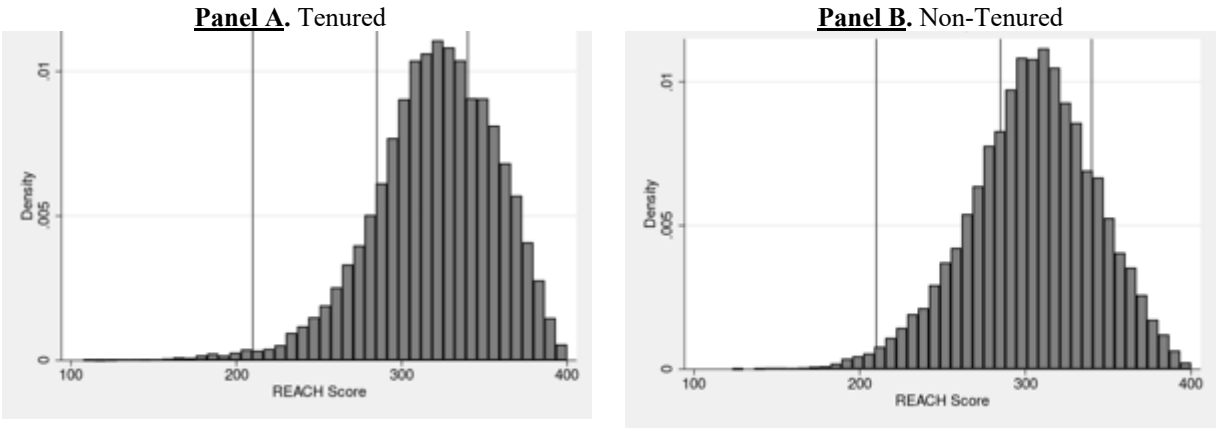


Panel H. Involuntary Exit (Year *t+1*), Non-Tenured



Notes. Each panel shows the exit rates for teachers with different REACH scores. Each point represents the average exit rate of teachers within a 7-point bin of the REACH score.

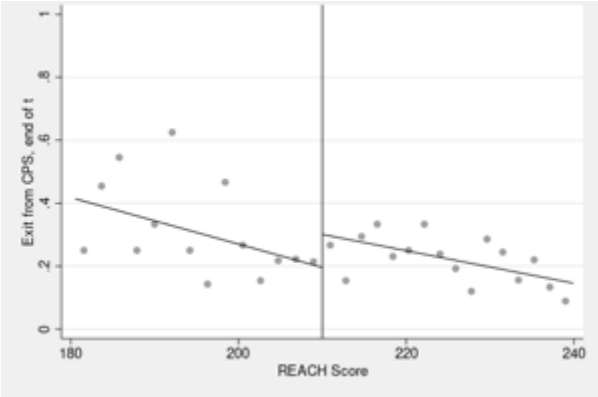
Figure 2. Distribution of REACH Score, by Tenure Status



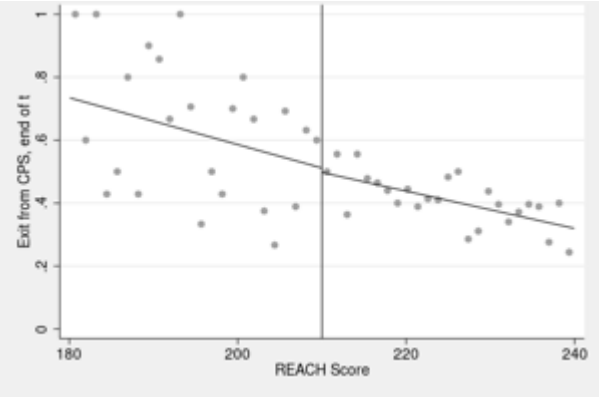
Notes. We tested for discontinuities at each of the three ratings thresholds and found no statistically significant discontinuities. For tenured teachers, the p-value from a McCrary (2008) test is 0.836 at the Unsatisfactory/Developing threshold; 0.537 at the Developing/Proficient threshold; and 0.279 at the Proficient/Excellent threshold. For non-tenured teachers, the p-value is 0.687 at the Unsatisfactory/Developing threshold; 0.683 at the Developing/Proficient threshold; and 0.778 at the Proficient/Excellent threshold.

Figure 3. Probability of Teacher Exit from CPS at the Unsatisfactory/Developing Threshold, by Tenure Status

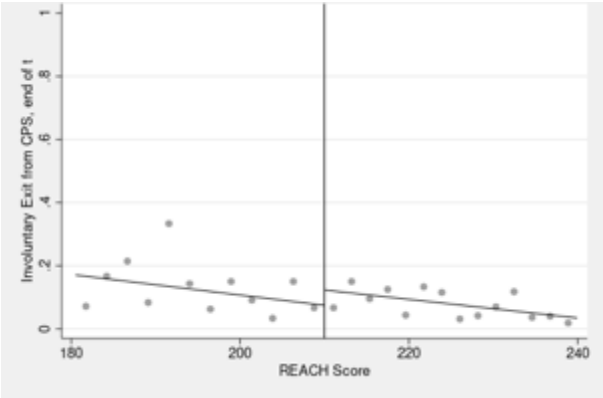
Panel A. Any Exit (Year *t*), Tenured



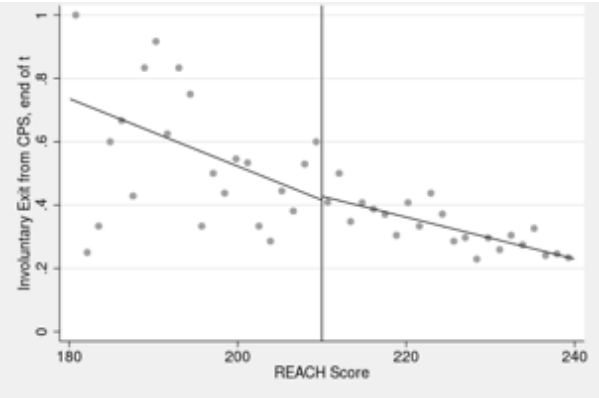
Panel B. Any Exit (Year *t*), Non-Tenured



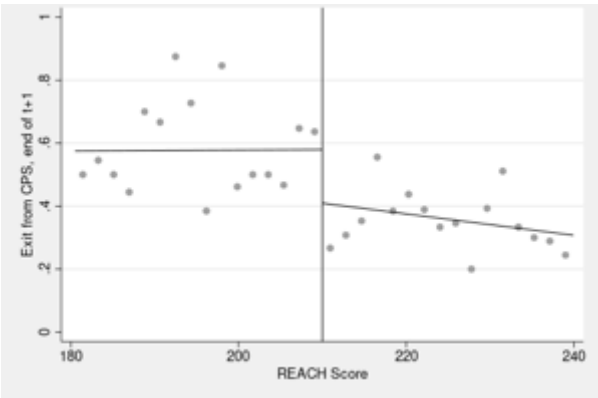
Panel C. Involuntary Exit (Year *t*), Tenured



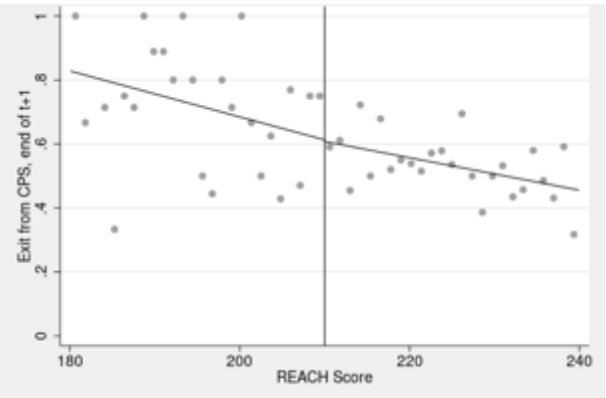
Panel D. Involuntary Exit (Year *t*), Non-Tenured



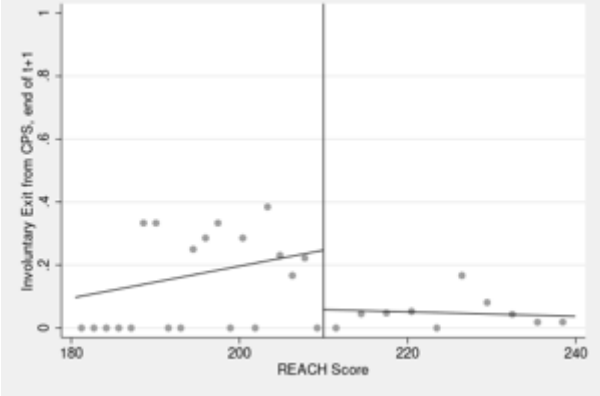
Panel E. Any Exit (Year *t+1*), Tenured



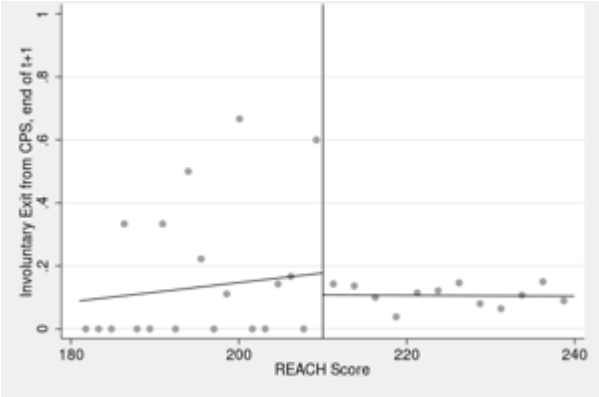
Panel F. Any Exit (Year *t+1*), Non-Tenured



Panel G. Involuntary Exit (Year *t+1*), Tenured



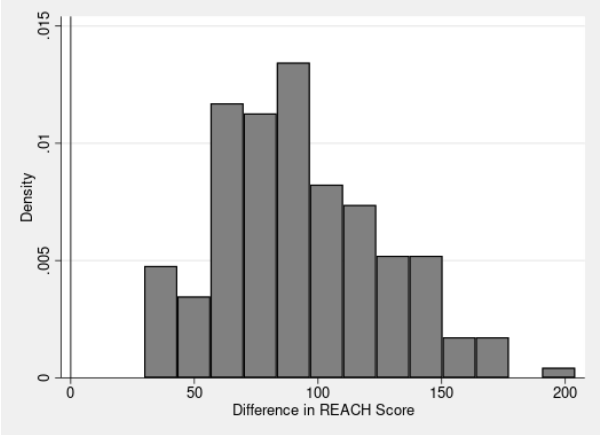
Panel H. Involuntary Exit (Year *t+1*), Non-Tenured



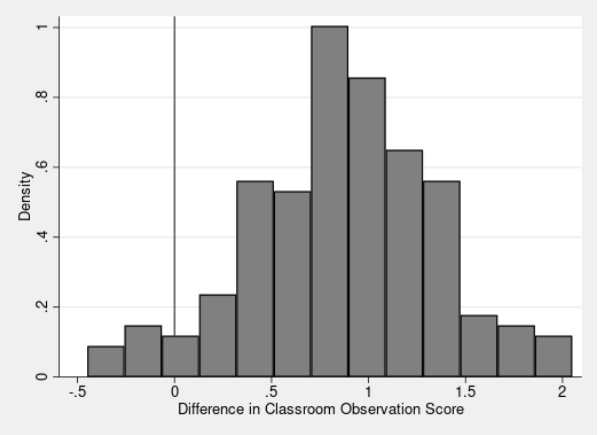
Notes. Each panel shows the exit rates for teachers with different REACH scores within 30 points of the Unsatisfactory/Developing threshold of 210 REACH score points. In each panel, the solid lines are local linear fits; dots are within bin averages. The number of bins is allowed to differ to the right and left of the cutoff and is selected using the mimicking variance evenly-spaced method (Calonico et al. 2017). The left-hand-side panels limit the sample to tenured teachers; the right-hand-side panels limit the sample to non-tenured teachers.

Figure 4. Distribution of the Difference in Performance Measures between the Average New-to-CPS Replacement Teachers and Unsatisfactory-Rated Teachers

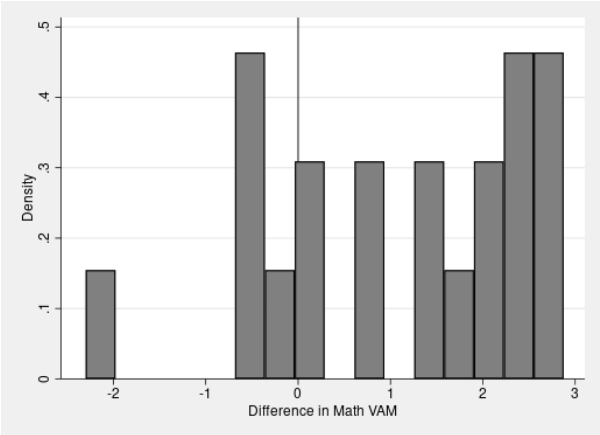
Panel A. Difference in REACH scores



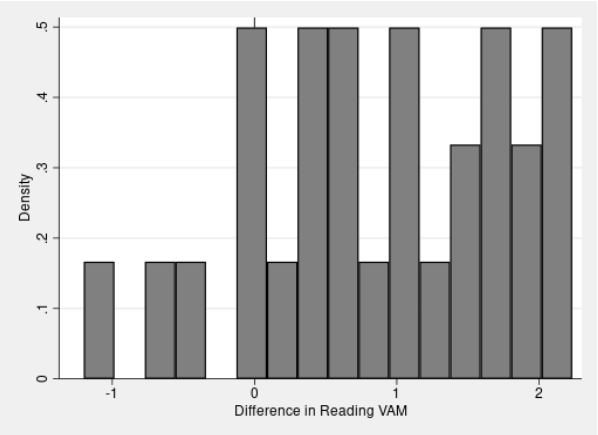
Panel B. Difference in classroom observation scores



Panel C. Difference in math VAM



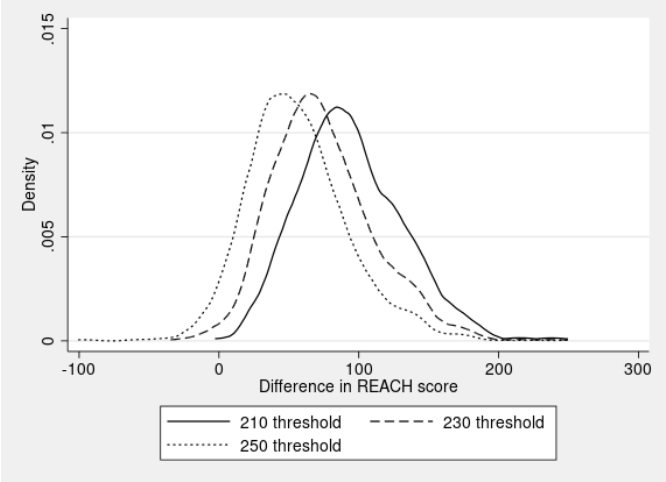
Panel D. Difference in reading VAM



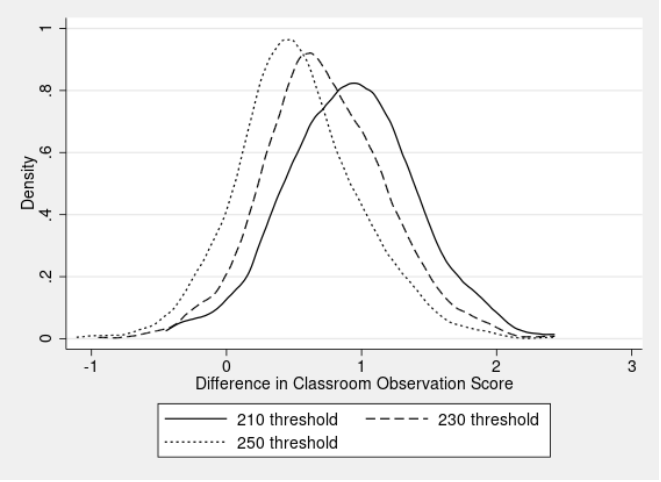
Notes. In each figure, a value below 0 indicates that the average new-to-school replacement teacher is lower performing on that metric than the Unsatisfactory-rated teacher. The number of cases where this occurs is n=1 of 355 for the REACH score, n=10 of 363 for the classroom observation score, n=11 of 42 for the math VAM, and n=11 of 47 for the reading VAM.

Figure 5. Distribution of the Difference in Performance Measures between New-to-CPS Replacement Teachers and Unsatisfactory-Rated Teachers, by REACH Score Thresholds

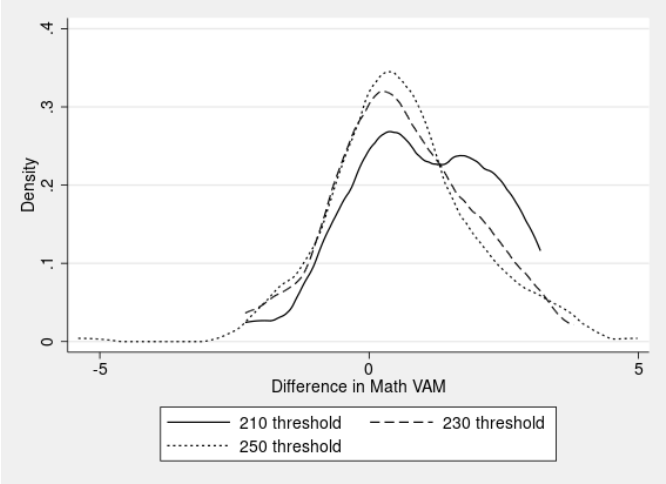
Panel A. Difference in REACH scores



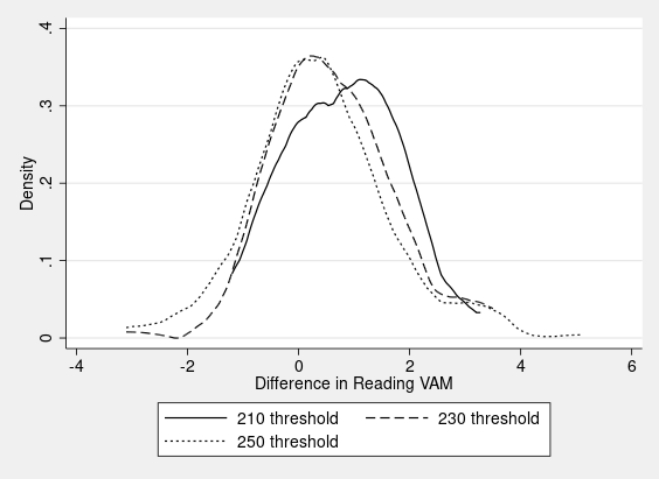
Panel B. Difference in classroom observation scores



Panel C. Difference in Math VAM



Panel D. Difference in Reading VAM



Notes. A value below 0 means that the average new-to-CPS replacement teacher is lower performing on that metric than the Unsatisfactory-rated teacher. For the 210 threshold (n=537 teachers below the threshold), the number of cases where this occurs is n=1 for the REACH score, n=10 for the classroom observation score, n=10 for the math VAM, and n=11 for the reading VAM. For the 230 threshold (n=1,217 teachers below the threshold), the number of cases where this occurs is n=8 for the REACH score, n=42 for the classroom observation score, n=33 for the math VAM, and n=34 for the reading VAM. For the 250 threshold (n=2,748 teachers below the threshold), the number of cases where this occurs is n=66 for the REACH score, n=207 for the classroom observation score, n=77 for the math VAM, and n=98 for the reading VAM.

Appendix Tables & Figures

Table A1. REACH System Teacher Performance Measures and Associated Weights

	Grades 3-8 in tested subject	Grades K-8 in non-tested subject/grade level	Grades 9-12
Individual value-added measures based on standardized test scores	20%	-	-
Student growth on district-developed assessments	10%	30%	30%
Classroom observation ratings	70%	70%	70%

Notes. Each cell provides the nominal weight assigned to a teacher performance measure used to construct a teacher's final REACH score upon which the final REACH evaluation ratings are based. District-developed assessments, which are written or hands-on assessments specifically designed for the grade and subject of the course, are administered and scored by teachers at the beginning and the end of the year. These assessments fulfill the state legislative requirement that all teachers should be evaluated in part based on student growth. Individual value-added measures are based on the NWEA-MAP. Classroom observation ratings are based on administrator observations of teacher practice using the Danielson Framework for Teaching. Weights for each of the performance measures have changed slightly throughout the implementation of REACH, but classroom observation ratings have always been the most heavily weighted component.

Table A2. Timing of REACH System Evaluation Cycle

	Year t-1	Year t	Year t+1
REACH data collected	X (only high-rated tenured teachers)	X (all teachers)	
Formal REACH rating awarded			X (October/November)
Labor market response to informal information		X (end of period t)	
Labor market response to formal rating			X (end of period t+1)

Notes. For tenured teachers who previously received Proficient or Excellent, REACH data is collected over a two-year period. For tenured teachers who previously received Unsatisfactory or Developing ratings and all non-tenured teachers, the data collection period is one year.

Table A3. Discontinuities in Teacher Characteristics at the Evaluation Rating Thresholds

Teacher characteristic	Unsatisfactory/ Developing Threshold	Developing/ Proficient Threshold	Proficient/ Excellent Threshold
P(teacher = black)	0.01	0.01	-0.02
	(0.07)	(0.02)	(0.02)
P(teacher = white)	0.00	-0.01	-0.02
	(0.07)	(0.02)	(0.02)
P(teacher = female)	-0.01	0.00	0.00
	(0.07)	(0.02)	(0.01)
P(teacher holds grad degree)	0.06	-0.01	0.00
	(0.07)	(0.02)	(0.02)
P(teacher = National Board)	0.00	-0.02***	0.01
	(0.02)	(0.01)	(0.01)
Birth year	-1.95	0.01	-0.53*
	(1.62)	(0.39)	(0.32)

Note. Each cell reports results from a separate nonparametric regression where the outcome is a specific teacher characteristic and the coefficient is the effect of being below the threshold. Regressions include only year fixed effects as controls and are restricted to a bandwidth of 25 points around the threshold. Robust standard errors are in parentheses. Coefficients are statistically significant at the *10%, **5% and ***1% levels.

Table A4. Nonparametric RD Estimates of the Impact of Proficient Evaluation Rating on Teacher Exit, by Tenure Status

	Any Exit (Year t)		Involuntary Exit (Year t)		Any Exit (Year t+1)		Involuntary Exit (Year t+1)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Tenured								
Proficient (relative to Developing)	.007 (.013)	.004 (.013)	.000 (.007)	-.001 (.007)	.033 (.020)	.030 (.020)	-.012 (.010)	-.012 (.010)
Counterfactual Mean	0.07	0.07	0.01	0.01	0.14	0.14	0.01	0.01
Bandwidth	32.5	31.3	23.7	23.4	26.1	25.5	16.7	16.6
N (left)	2,786	2,704	2,290	2,264	3,462	2,407	1,268	1,263
N (right)	7,212	6,865	4,870	4,792	5,863	5,341	2,227	2,211
Panel B: Non-tenured								
Proficient (relative to Developing)	.019 (.016)	.020 (.016)	.014 (.012)	.013 (.012)	.032* (.019)	.034* (.019)	.001 (.012)	.001 (.012)
Counterfactual Mean	0.12	0.12	0.03	0.03	0.22	0.22	0.03	0.03
Bandwidth	28.3	28.3	25.1	24.6	32.9	32.9	21.8	21.6
N (left)	3,462	3,462	3,207	3,171	3,810	3,810	2,111	2,100
N (right)	5,863	5,863	5,176	5,066	6,821	6,821	3,348	3,319
Year FE	X	X	X	X	X	X	X	X
Teacher Xs		X		X		X		X

Notes. Each column (within a panel) is a separate regression. Coefficients from nonparametric regression discontinuity (RD) reported with robust standard errors (clustered at the school level). All regressions include controls for the linear running variable – a teacher's final REACH score (from year *t*). *Teacher Xs* include controls for teacher gender, race/ethnicity, educational attainment, National Board Certification, and birth year. Coefficients are statistically significant at the *10%, **5% and ***1% levels.

Table A5. Nonparametric RD Estimates of the Impact of Excellent Evaluation Rating on Teacher Exit, by Tenure Status

	Any Exit (Year t)		Involuntary Exit (Year t)		Any Exit (Year t+1)		Involuntary Exit (Year t+1)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Tenured								
Excellent (relative to Proficient)	-.015 (.011)	-.011 (.011)	.002 (.004)	.003 (.004)	-.012 (.016)	-.009 (.015)	-.001 (.008)	-.009 (.008)
Counterfactual Mean	0.05	0.05	0.01	0.01	0.11	0.11	0.01	0.01
Bandwidth	14.1	15.5	20.8	20.3	15.2	17.0	19.4	20.3
N (left)	3,540	3,966	5,441	5,311	3,899	4,379	2,803	2,960
N (right)	3,265	3,570	4,578	4,492	3,498	3,873	2,183	2,256
Panel B: Non-tenured								
Excellent (relative to Proficient)	-.052** (.022)	-.054** (.022)	-.003 (.008)	.002 (.008)	-.069** (.031)	-.074** (.032)	.019 (.013)	.019 (.012)
Counterfactual Mean	0.09	0.09	0.01	0.01	0.18	0.18	0.02	0.02
Bandwidth	12.0	11.6	16.2	16.5	10.9	10.3	12.5	13.1
N (left)	1,818	1,743	2,543	2,615	1,632	1,517	1,299	1,383
N (right)	1,479	1,440	1,851	1,877	1,377	1,314	1,014	1,057
Year FE	X	X	X	X	X	X	X	X
Teacher Xs		X		X		X		X

Notes. Each column (within a panel) is a separate regression. Coefficients from nonparametric regression discontinuity (RD) reported with robust standard errors (clustered at the school level). All regressions include controls for the linear running variable – a teacher's final REACH score (from year *t*). *Teacher Xs* include controls for teacher gender, race/ethnicity, educational attainment, National Board Certification, and birth year. Coefficients are statistically significant at the *10%, **5% and ***1% levels.

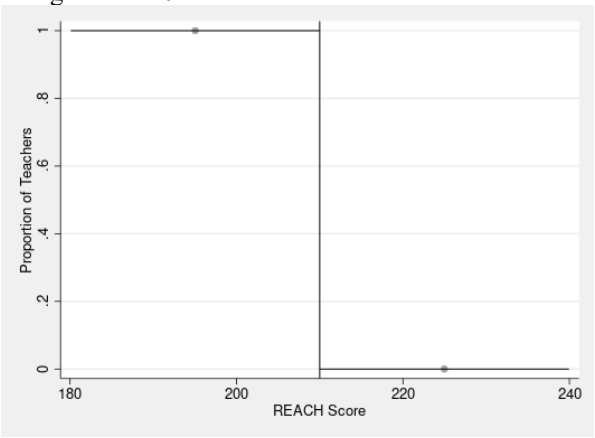
Table A6. Performance Comparison of Unsatisfactory-Rated Teachers to All Replacement Teachers in the Same School

	New to School (Year t+1)			New to School (Year t+2)		
	Lowest Rated	Average Rated	Highest Rated	Lowest Rated	Average Rated	Highest Rated
Panel A: *Unsatisfactory Teachers*						
REACH Score	57.08	96.88	132.00	58.18	99.79	136.00
	(43.88)	(35.64)	(41.92)	(43.42)	(34.86)	(38.27)
Classroom Observation	0.43	0.96	1.40	0.48	1.00	1.43
	(0.59)	(0.45)	(0.50)	(0.59)	(0.44)	(0.46)
VAM (Math)	0.45	0.87	1.35	0.34	0.69	1.03
	(1.33)	(1.28)	(1.42)	(1.47)	(1.37)	(1.44)
VAM (Reading)	0.58	1.01	1.47	0.26	0.72	1.17
	(1.51)	(1.40)	(1.56)	(1.57)	(1.37)	(1.40)
Panel B: *Unsatisfactory Teachers Who Remain in t+1*						
REACH Score	55.27	94.12	127.25	59.66	98.82	133.77
	(41.96)	(34.11)	(42.21)	(41.92)	(31.36)	(32.66)
Classroom Observation	0.37	0.90	1.31	0.50	0.98	1.37
	(0.59)	(0.45)	(0.52)	(0.57)	(0.42)	(0.41)
VAM (Math)	0.39	0.78	1.24	0.18	0.56	0.94
	(1.38)	(1.34)	(1.50)	(1.68)	(1.60)	(1.69)
VAM (Reading)	0.72	1.17	1.66	0.29	0.73	1.15
	(1.30)	(1.12)	(1.25)	(1.63)	(1.35)	(1.37)

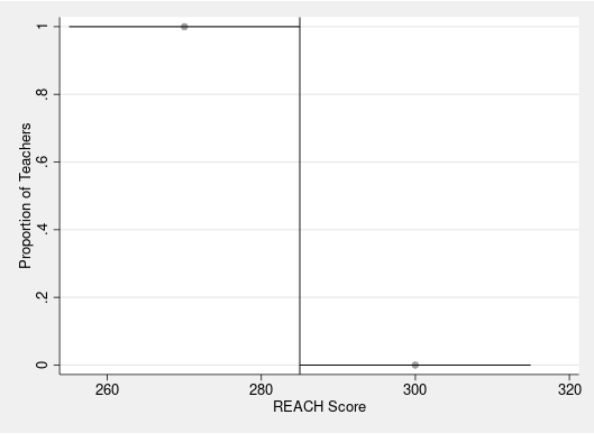
Notes. Each cell reports the mean (standard deviation) difference in teacher performance (by performance measure) between CPS teachers rated Unsatisfactory (in school year t) and teachers who are new to the same school in subsequent school years (either year t+1 or year t+2). In all cells, positive values indicate that the new-to-school replacement teachers are higher performing than the Unsatisfactory rated teachers.

Figure A1. Probability of Receiving a Formal REACH Evaluation Rating Given the REACH Score

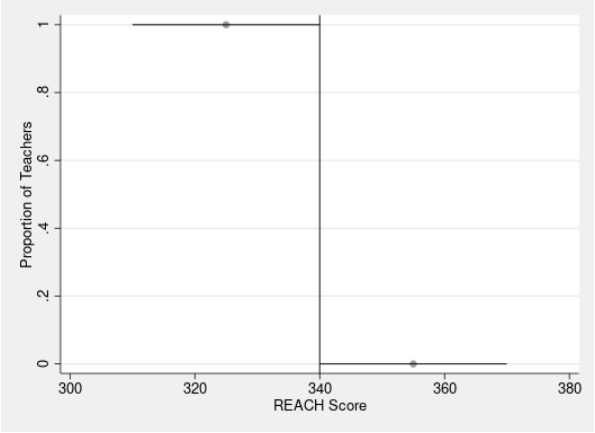
Panel A. Probability of receiving an Unsatisfactory rating at the 210 threshold



Panel B. Probability of receiving a Developing rating at the 285 threshold



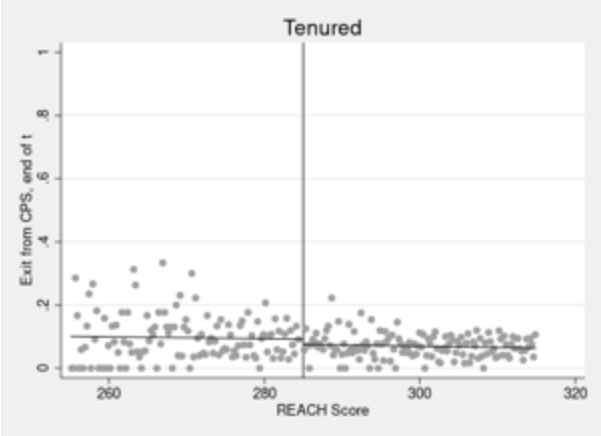
Panel C. Probability of receiving a Proficient rating at the 340 threshold



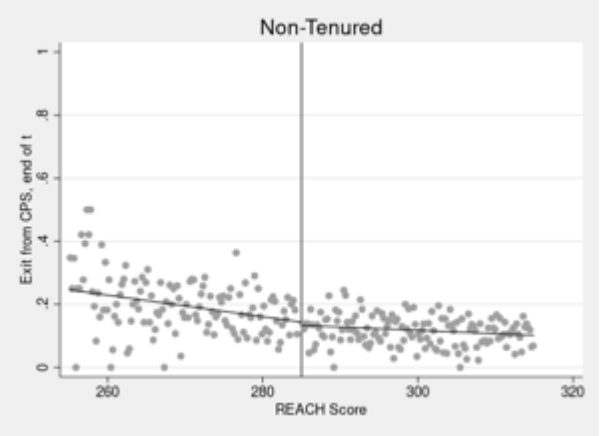
Notes. Each panel shows the share of teachers with different REACH scores who received a given rating within 30 points of a given evaluation rating threshold.

Figure A2. Probability of Teacher Exit from CPS at the Developing/Proficient Threshold, by Tenure Status

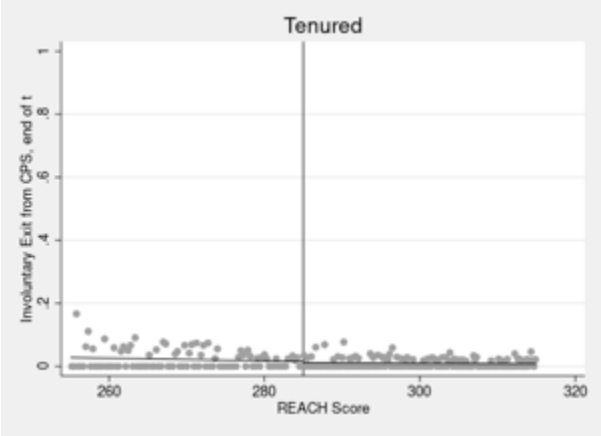
Panel A. Any Exit (Year *t*), Tenured



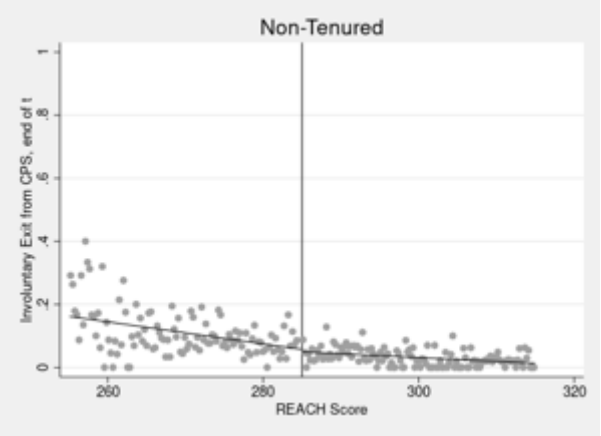
Panel B. Any Exit (Year *t*), Non-Tenured



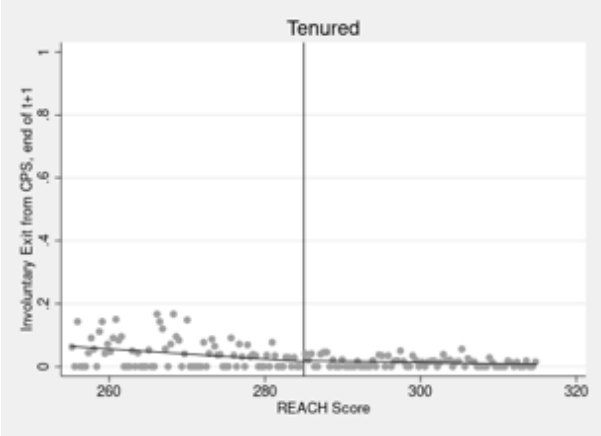
Panel C. Involuntary Exit (Year *t*), Tenured



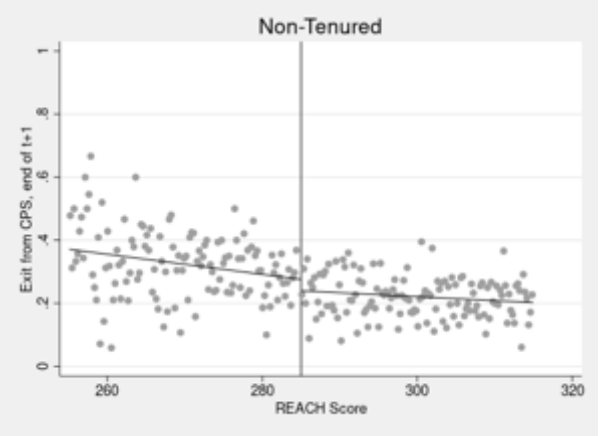
Panel D. Involuntary Exit (Year *t*), Non-Tenured



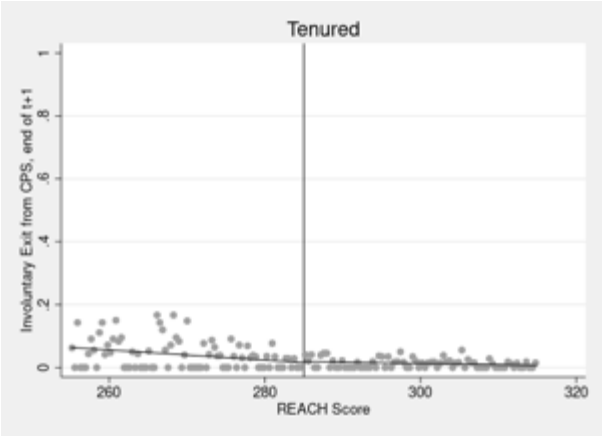
Panel E. Any Exit (Year *t+1*), Tenured



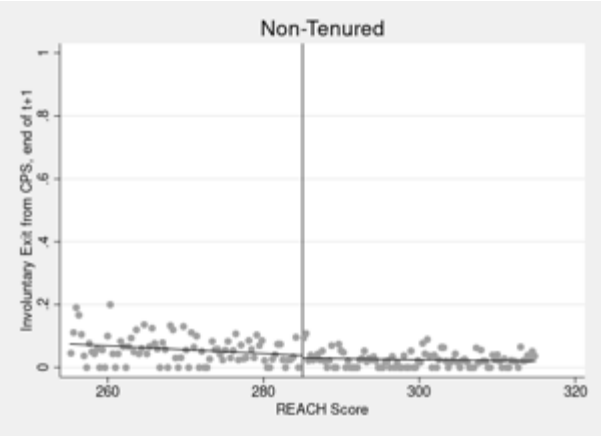
Panel F. Any Exit (Year *t+1*), Non-Tenured



Panel G. Involuntary Exit (Year *t+1*), Tenured



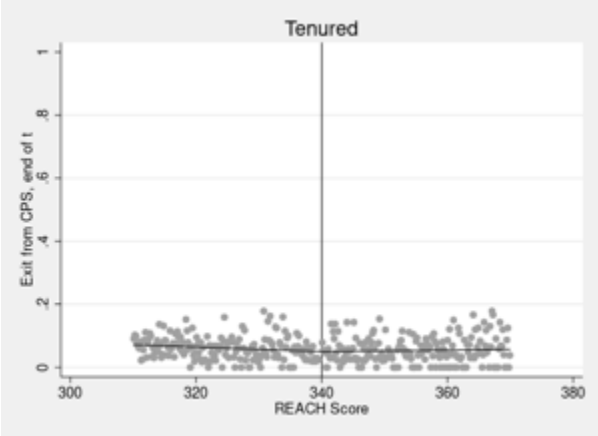
Panel H. Involuntary Exit (Year *t+1*), Non-Tenured



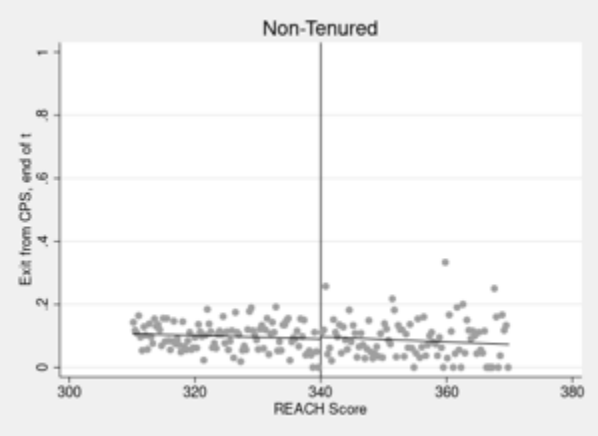
Notes. Each panel shows the exit rates for teachers with different REACH scores within 30 points of the Developing/Proficient threshold of 285. In each panel, the solid lines are local linear fits; dots are within bin averages. The number of bins is allowed to differ to the right and left of the cutoff and is selected using the mimicking variance evenly spaced method (Calonico et al. 2017). The left-hand-side panels limit the sample to tenured teachers; the right-hand-side panels limit the sample to non-tenured teachers.

Figure A3. Probability of Teacher Exit from CPS at the Proficient/Excellent Threshold, by Tenure Status

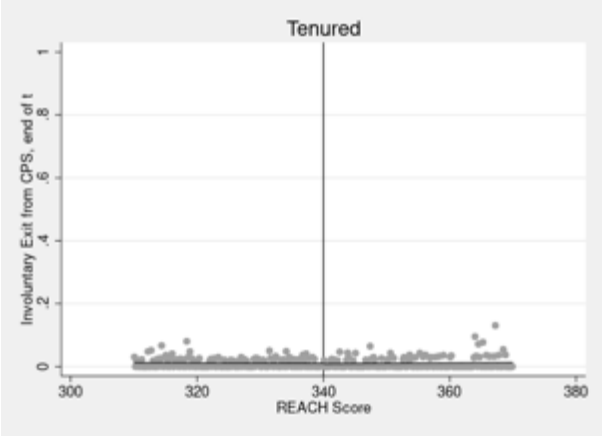
Panel A. Any Exit (Year *t*), Tenured



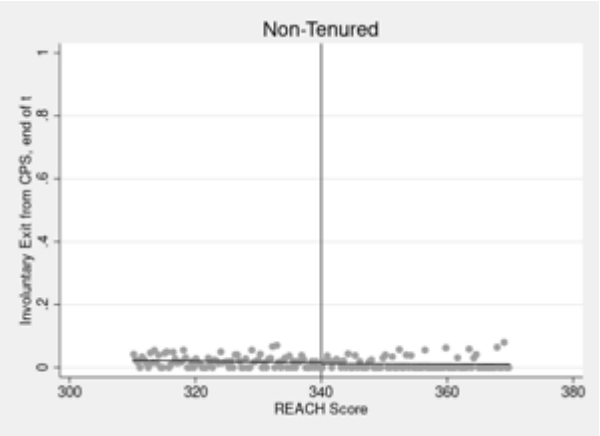
Panel B. Any Exit (Year *t*), Non-Tenured



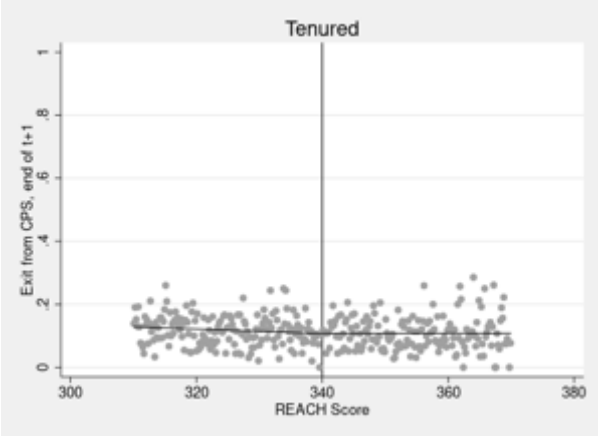
Panel C. Involuntary Exit (Year *t*), Tenured



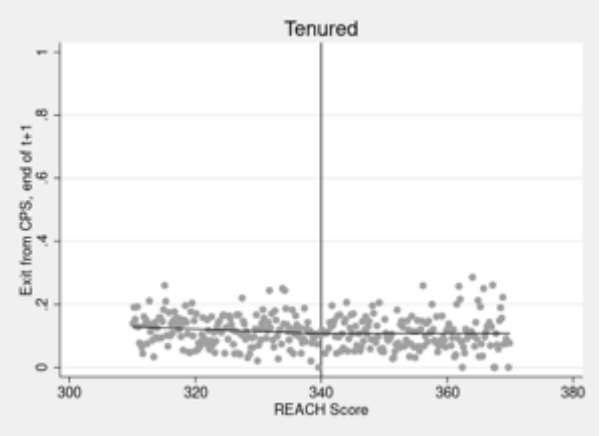
Panel D. Involuntary Exit (Year *t*), Non-Tenured



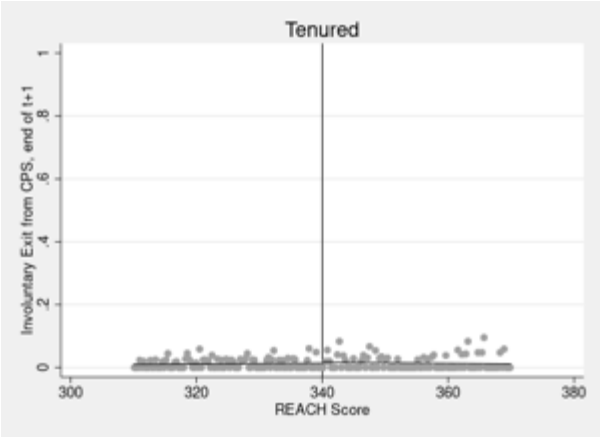
Panel E. Any Exit (Year *t+1*), Tenured



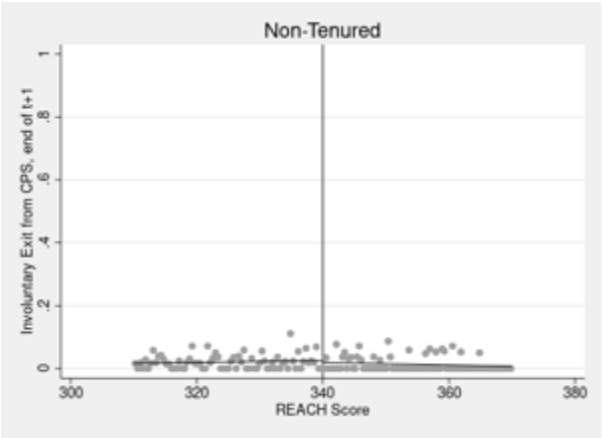
Panel F. Any Exit (Year *t+1*), Non-Tenured



Panel G. Involuntary Exit (Year *t+1*), Tenured



Panel H. Involuntary Exit (Year *t+1*), Non-Tenured



Notes. Each panel shows the exit rates for teachers with different REACH scores within 30 points of the Proficient/Excellent threshold of 340. In each panel, the solid lines are local linear fits; dots are within bin averages. The number of bins is allowed to differ to the right and left of the cutoff and is selected using the mimicking variance evenly spaced method (Calonico et al. 2017). The left-hand-side panels limit the sample to tenured teachers; the right-hand-side panels limit the sample to non-tenured teachers.