NBER WORKING PAPER SERIES

HOW DOES SCHOOL ACCOUNTABILITY AFFECT TEACHERS? EVIDENCE FROM NEW YORK CITY

Rebecca Dizon-Ross

Working Paper 24658
http://www.nber.org/papers/w24658

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2018

I want to thank the New York City Department of Education for providing me with the data, and Dominique West and Marsha Modeste for answering all of my data questions. I am very grateful to Ran Abramitzky, Pascaline Dupas, Caroline Hoxby, and Seema Jayachandran for help and guidance, and to Marinho Bertanha, Elise Dizon-Ross, David Figlio, Anil Jain, Jonah Rockoff, Fabiana Silva, Jenny Ying, and numerous participants at the Stanford Applied Lunch for helpful comments. Christine Cai provided excellent research assistance. All errors are my own. I appreciate the generous support of the Shultz Graduate Student Fellowship in Economic Policy, the endowment in memory of B.F. Haley and E.S. Shaw, the Spencer Foundation, and the National Science Foundation. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

How Does School Accountability Affect Teachers? Evidence from New York City
Rebecca Dizon-Ross
NBER Working Paper No. 24658
May 2018
JEL No. I2

ABSTRACT

Does holding schools accountable for student performance cause good teachers to leave low-performing schools? Using data from New York City, which assigns accountability grades to schools based on student achievement, I perform a regression discontinuity analysis and find evidence of the opposite effect. At the bottom end of the school grade distribution, I find that a lower accountability grade decreases teacher turnover and increases joining teachers' quality. A likely channel is that accountability pressures induce increases in principal effort at lower-graded schools, especially among high-quality principals, and teachers value these changes. In contrast, at the top end of the school grade distribution, where accountability pressures are lower, low accountability grades may negatively impact joining teachers' quality.

Rebecca Dizon-Ross
Booth School of Business
University of Chicago
5807 South Woodlawn Avenue
Chicago, IL 60637
and NBER
rdr@chicagobooth.edu

1 Introduction

Since the mid-1990's, school accountability systems have become a central focus of education reform in the United States. Even before the No Child Left Behind Act (NCLB) made accountability mandatory across the U.S. in 2001, many states and districts had already instituted some form of accountability.

Policymakers and observers often worry that these systems, which attempt to hold schools accountable for student performance, could make it difficult for low-performing schools to attract and retain good teachers. Evidence from surveys of teachers suggests that good teachers may want to avoid the stress, restricted autonomy, and emphasis on "teaching to the test" that they think accountability brings to low-performing schools (Jones et al., 1999; Kirtley, 2012). Since high-quality teachers improve students' long-run educational attainment and earnings more than low-quality teachers (Chetty et al., 2014; Kane et al., 2008), this means that accountability systems could have lasting, negative implications for students at low-performing schools. Indeed, some have suggested that poor accountability ratings could start low-performing schools down a negative quality spiral where high-quality teachers leave, causing the best students to leave, thus causing more teachers to leave, and so forth.

However, from a theoretical perspective, the effect of low school accountability ratings on teacher quality is actually ambiguous. For example, teachers could *prefer* to teach in lower-rated schools. Although that may sound counterintuitive, one potential reason is school improvement: low accountability ratings are designed to improve school performance, and a large literature shows that they do (Carnoy and Loeb, 2002; Chiang, 2009; Figlio and Rouse, 2006; Hanushek and Raymond, 2005; West and Peterson, 2006; Rockoff and Turner, 2010). Teachers may prefer to teach in schools where achievement is improving, perhaps because they value achievement or because the process of improvement is satisfying. Thus, it is ultimately an empirical question whether the impact of accountability pressures on teachers is positive or negative.

This paper exploits the introduction of an accountability system in the New York City Department of Education (hereafter: NYCDOE) to provide new evidence on this issue. In November, 2007, the NYCDOE launched a comprehensive accountability system which assigned schools letter grades for school performance. The grades were based on continuous performance metrics, with determination of the actual grade based on strict thresholds. This allows for the use of a regression discontinuity analysis to estimate the effect of the reform, as previously shown by Rockoff and Turner (2010). While these authors focused on the within-year impacts of receiving a low grade on student performance (finding large positive improvements), here I focus on the impacts on the teacher labor market.

Using data from the first two years of NYCDOE's accountability system, I find evidence that accountability pressures can in some cases *improve* teacher retention and recruitment. At the bottom end of the school grade distribution (i.e., the C/D and D/F thresholds), where the NYCDOE accountability sanctions have more "bite," receipt of a lower school accountability grade, which occurs early in the school year, decreases teacher turnover at the end of the year by three percentage points. This is a large effect, representing roughly 20% of baseline turnover. This pattern is robust across several different specifications: I consistently reject the null of no effect and, a fortiori, the conventional wisdom that low accountability scores might lead to a sizeable increase in teacher turnover. The decrease in turnover likely benefits students in low-graded schools, as turnover has been shown to decrease student achievement (Ronfeldt et al., 2013).

I next examine the sorting implications of accountability grades, and again find evidence that lower accountability grades help the low-performing schools at the bottom end of the grade distribution. The evidence suggests that joiners to lower-graded schools have higher value-added than the joiners to higher-graded schools, but there is no difference in the value-added of leavers between lower-graded and higher-graded schools.

There are two main hypotheses that could explain these effects. The first hypothesis is that receiving a lower school grade increases the attractiveness of the jobs at the school (the *job desirability* hypothesis). One potential reason is school improvement: Rockoff and Turner (2010) show that the lower-rated schools in the NYCDOE improved their performance within the same year the grade was received. Teachers may prefer to teach in schools where achievement is improving, because, as discussed above, they value achievement per se, appreciate the changes that enable the performance improvements, or expect that it will lead to higher accountability grades in the future. Another potential channel for *job desirability* is that, induced by accountability pressure, principals at lower-graded schools put more effort into making the schools better places for teachers to work, or into attracting and retaining high-quality teachers. The second hypothesis is that receiving a lower school grade attaches a negative stigma to the teachers at the school, reducing their perceived value to potential employers who view low grades as a signal of low unobservable teacher quality (the *stigma* hypothesis). Note that this is a rational hypothesis.[1]

I argue that the *job desirability* hypothesis matches the data better than the *stigma* hypothesis. The fact that lower-graded schools have higher-quality joiners than higher-graded schools is more consistent with increased job desirability. Also suggestive is the fact that the effect of lower grades on turnover is primarily driven by a decrease in out-of-district departures. Data on the transfer applications submitted by teachers also show that teachers who applied to transfer out of lower-graded schools were no less likely to be able to transfer out, and that the number of teachers who applied to transfer to lower-graded schools did not fall.

The results of surveys conducted with teachers provide additional evidence for the potential mechanisms underlying *job desirability*. At the bottom end of the grade distribution, teachers at schools that received lower accountability grades at the beginning of the year gave their principals higher leadership ratings at the end of the year than teachers at higher-graded schools, agreeing more strongly with statements such as "the principal is an effective manager who makes the school run smoothly" and "I feel strongly supported by my principal" (note that the results are significant at the 10% level). Since principal leadership ratings also correlate with lower turnover, these effects suggest that one channel for the positive impact of lower grades may be principals at low-graded schools respond to accountability pressures by making management changes that appeal to teachers.

Given the important role for principals suggested by these results, I also look at heterogeneity in

[1] This type of stigma is distinct from the social stigma that the existing accountability literature has often deemed responsible for spurring test score improvements in response to low accountability grades (Figlio and Rouse, 2006; West and Peterson, 2006). In this paper's categorization, social stigma would in fact impact job desirability. The direct effect would be negative, but the indirect (and thus net) effect could be positive if it spurs test score improvements that teachers value; this is thus a potential channel for job desirability to explain this paper's findings.

the RD impacts of low grades by baseline measures of principal quality. I find that the decreases in turnover at lower-graded schools are driven primarily by schools with higher-quality principals (as measured by baseline principal leadership ratings). This suggests that principal capacity to transform accountability pressures into positive changes for the school may be a key ingredient that enables low accountability grades to have positive labor market effects.

Thus, the results suggest that the accountability system benefited low-performing schools at the bottom end of the grade distribution through two labor market channels: decreased turnover and increased teacher quality. These effects appear to be driven by the low-graded schools becoming more attractive to teachers. However, at the top end of the school grade distribution (the A/B and B/C thresholds), the results differ. Here, I find no evidence of positive effects of receiving a lower grade, and some suggestive evidence of negative impacts on the quality of the joiners relative to the leavers and on teacher survey responses about their principals' leadership.

The difference in the results at the top and bottom ends of the grade distribution (i.e., the fact that accountability seems to benefit lower-rated schools at the C/D and D/F thresholds while hurting them at the A/B and B/C thresholds) likely reflects the fact that accountability pressures are higher at the C/D and D/F thresholds, and so only motivated positive changes there (Rockoff and Turner, 2010).[2] Are there other ingredients besides high stakes that help us predict when accountability systems will positively impact the teacher labor market? The results on mechanisms suggest that positive impacts may be more likely in settings where principals are good leaders and have the latitude to implement positive changes, but of course such discussion is still speculative at this point.

This paper contributes to the literature on determinants of teacher mobility and sorting (Scafidi et al., 2007; Falch and Rønning, 2007; Imazeki, 2005; Hanushek et al., 2004; Hendricks, 2014; Dolton and von der Klaauw, 1995; Hensvik, 2012; Venhorst et al., 2011; Jackson, 2009, 2012). Within the more specific literature examining how accountability affects teachers, the paper addresses two of the primary challenges. The first challenge is identification: accountability reforms

[2] Specifically, the explanation is that teachers prefer schools that (1) have improved in some way (e.g., environment, performance, principal effort) in response to accountability pressures, and (2) have a higher nominal accountability grade. At the top end of the grade distribution, accountability does not bind, and so (1) plays no role while (2) dominates. At the bottom end of the grade distribution, low-performing schools do improve in response to accountability, and so (1) dominates.

are often instituted simultaneously with many other reforms that also affect teachers, making it difficult to cleanly identify accountability's effects. As with all regression discontinuity designs, the analysis used in this paper focuses on schools right next to the grade thresholds, and thus holds fixed the effects of concurrent reforms. This builds on some earlier papers in the literature that used difference-in-differences strategies which could be more subject to these identification concerns (Clotfelter et al., 2004; Boyd et al., 2008), with Feng et al. (2010) and Gjefsen and Gunnes (2016) more recent exceptions. Feng et al. (2010) use unexpected changes to the school accountability grading system in Florida that exogenously "shocked" some schools' grades, while Gjefsen and Gunnes (2016) uses triple difference specifications.[3]

The second challenge is finding good data on teacher quality, and specifically, on teachers' contributions to student learning, or their "value-added." Having a good measure of teacher quality is important for understanding the implications of accountability: high turnover could either reflect high-quality teachers leaving or low-quality teachers being pushed out. Value-added is widely regarded as unmatched as a measure of teacher quality, with, for example, important predictive power over students' long-run outcomes (Chetty et al., 2014). Unfortunately, value-added estimation has extensive data requirements, and so most of the existing literature (Boyd et al., 2008; Clotfelter et al., 2004; Gjefsen and Gunnes, 2016) could only use other teacher characteristics, which often do not proxy well for value-added (see, e.g., Hanushek et al. (2010) and Rivkin et al. (2005)).[4]

This paper makes several new contributions that go beyond the results contributed by Feng et al. (2010), which also make use of value-added data to examine a similar question. First, my analysis speaks to how accountability affects teacher quality in a more comprehensive manner, by incorporating value-added data on joining teachers, not just leaving teachers. This is important because, as the results here show, there can be asymmetric effects for the two groups, and it is the comparison of the two effects that determines the overall impact. Second, this is the first paper to unpack the *mechanisms* behind accountability's effects on teachers by trying to disentangle the two main factors that affect teacher mobility: changes in choice sets vs. changes in job desirability.

[3] On a related topic, Shirrell (2016) uses RD to look at the effect of "subgroup-specific" accountability (i.e., whether a school was accountable for the performance of specific racial/ethnic student subgroups or not) on teacher turnover.

[4] In related work, Li (2011) uses data on *principal* value-added to examine how accountability affects principals.

This helps us learn not just about accountability's channels but also about teacher preferences and mobility decisions more broadly.

In contrast to my findings, previous literature looking at similar phenomena largely concludes that accountability pressures hurt low-performing schools by accelerating turnover (Feng et al., 2010; Clotfelter et al., 2004; Gjefsen and Gunnes, 2016), more so at the bottom of the grade distribution in Feng et al. (2010). Section 7 discusses potential reasons for the different results.

The remainder of the paper proceeds as follows: Section 2 describes the institutional background. Section 3 describes the data and empirical strategy. Section 4 presents the main results, while Section 5 discusses potential mechanisms for the results. Section 6 examines robustness. In Section 7, I discuss my results in the context of the overall literature. Section 8 concludes.

2 Background

2.1 The NYCDOE Accountability System

I now review the key features of the NYCDOE accountability system, much of which was previously described in Rockoff and Turner (2010). The NYCDOE launched its current accountability system in November of 2007. Under the system, schools receive progress reports with letter grades meant to capture school performance relative to peer schools. The progress report also contains the school's NCLB status, and the score from a school's Quality Review, a 2-3 day qualitative evaluation. The NYCDOE links the letter grades with rewards and sanctions, and makes the reports publicly available in an effort to incentivize low-performing schools to improve their performance.

The letter grade is based on a numeric score. For elementary and middle schools (the focus of this study), the score reflects three measures: student progress, student performance, and school environment. Student progress represents 60% of the score and measures year-to-year changes in student scores on the New York State standardized tests in Mathematics and English Language Arts (ELA). Student performance (25% of the score) captures the *level* of test scores. School environment (15% of the score) reflects attendance and parent, student, and teacher surveys results.

School scores are calculated as a weighted average of the school's "city horizon score" (1/3 weight), which compares the school to all others of the same school type (i.e., that serve the same grades), and its "peer horizon score" (2/3), which compares it to a peer group of up to 40 similar schools.[5] The overall pre-additional-credit score, which ranges from 0 to 100, is then calculated as the weighted average of the scores for each grading measure. Schools can also earn additional credit if their "high-need" students make "exemplary gains" (i.e., improve their performance by at least one-half of a proficiency level in ELA or Math). The credit is added to the school's pre-additional-credit score to determine the final score.

Thresholds for letter grade assignment are determined based upon the distribution within school type of pre-additional-credit scores. For example, in the first year of the program, the NYCDOE set the threshold for receipt of an A, B, C, and D at the 85th, 45th, 15th, and 5th percentiles of pre-additional-credit scores, respectively. Grades are then determined by comparing each school's score to the thresholds, with the thresholds strict (see Appendix Figure A1).

Consequences of grades

The NYCDOE links the letter grades with rewards and sanctions. Quoting the guidelines, "schools that are given an overall grade of A receive financial rewards, unless they score poorly on the Quality Review. Schools that receive an overall grade of D or F are subject to school improvement measures and target setting and, if no progress is made over time, possible leadership change, restructuring, or closure. The same is true for schools receiving a C for three years in a row. Over time, school organizations receiving an overall grade of F are likely to be closed. Ultimately, schools are accountable for making progress and receiving an overall grade of A, B, or C" (NYCDOE website, 2010).[6] The sanctions associated with receiving low accountability grades (i.e., D or F) are significant. For example, after receiving the first report cards in November 2007, the NYCDOE told five F schools in December that they would be closed immediately or phased out at the end of the school year. At the top end of the grade distribution, higher letter grades are associated with modest rewards. Principals

[5] To calculate the peer horizon score, the NYCDOE assigns each school a peer index based on student demographics (elementary and K-8 schools) or past test scores of current students (middle schools). They then sort schools by peer indices within school types to form peer groups, which consist of the 20 schools above and below a given school.

[6] The NYCDOE was unable to provide specifics about exactly what was entailed by the school improvement measures, except to confirm that the measures did not involve any additional resources being given to the schools, and that there is no record of any formal guidelines for the process.

of schools that had a score among the top 20% of schools and that received a Well Developed or Proficient quality review rating were eligible for bonuses of \$7,000 to \$25,000. Schools receiving an A and a Well Developed quality review rating received roughly \$33 per student in extra funds, to be used at the principal's discretion. Finally, schools receiving an A or B grade and a Well Developed or Proficient quality review rating received \$1,500 to \$3,000 per student that transferred in from an F school or a school not in good standing under NCLB.

Besides these consequences, there were no other financial ramifications of school grades; for example, D and F schools did not receive any additional funding, resources, or staff for school improvement.

2.2 The NYCDOE Teacher Transfer System

Since 2005, NYCDOE's staffing system has been an open market system built around the principle of "mutual consent": schools post openings, teachers apply, principals choose which teachers to hire, and then teachers decide whether to accept the job offers. This means that the effects estimated in this paper are the effects on equilibrium matches within a market-based system.

2.3 Differences Between NYCDOE and Other U.S. Contexts

The NYCDOE system has lower paperwork requirements for failing schools than some other accountability systems, such as Florida, where failing schools are required to complete regular, extensive reports (Rouse et al., 2013). The base NYCDOE system also does not link progress reports with teacher-level incentives, whereas North Carolina's and Florida's systems include teacher performance bonuses.[7] Teachers unions are much stronger in New York than Florida or North Carolina, which could partially explain some of the differences in accountability program design. Finally, some accountability systems give failing schools additional funding for school improvement, but that was not the case in the NYCDOE.

[7] There was a pilot schoolwide bonus program instituted in NYCDOE that did tie progress reports to teacher bonuses, but it was a randomized pilot conducted with a subsample of schools and was not part of the core program instituted for all schools. As discussed in Section 3.4, this program does not affect the study results. In North Carolina, teacher bonuses were guaranteed, whereas in Florida, schools received payments that could be used either for bonuses or other school improvement, at the school's discretion (Peterson, 2006).

3 Data and Empirical Strategy

3.1 Data

I use data from several sources within the NYCDOE. The accountability data come from publicly available files downloaded from the NYCDOE website. The data contain each school's accountability score and components, as well as NCLB status, quality review rating, and school identifiers. The data are available for the 2007-08 through 2011-12 school years (where the school year given is the year in which the accountability grade was released; report cards are released in fall of the school year and depend on performance results from the previous year.)

The second data source is demographic and exam performance data at the student level, provided by the NYCDOE and covering the 1998-99 through 2008-09 school years. The demographic data include gender, ethnicity, free-lunch, and special-education status. The exam performance data include student scores on Mathematics and ELA tests administered statewide in 4th and 8th grade, and citywide in the 3rd, 5th, 6th, and 7th grades.

The majority of the teacher data come from the NYCDOE HR/payroll system and contain teacher experience and education, and school- and grade-level identifiers, from the 1999-2000 through the 2009-2010 school years. This is my primary source of data on teacher turnover and mobility. Based on guidance from the NYCDOE, to calculate turnover, I define a teacher as having left a school if she leaves between May of one school year and November of the subsequent school year, since the (rare) midyear departures tend to reflect emergencies (e.g., sickness, birth) and would increase noise, but I show robustness to this definition.

I combine these data with annual data from the open market transfer system showing which teachers submitted transfer applications, with a tag for which applicants successfully transferred. Data are available from the 2006-2007 through 2009-2010 schoolyears. Although the HR/payroll system allows me to calculate transfers, the open market data is useful for understanding (a) which teachers *want* to transfer even if not successful, and (b) how many transfer applications different schools receive. However, the data have some weaknesses. First, although I see every application submitted by the 60% of teachers who did not ultimately transfer, for the 40% of applicants who do transfer, I only see the application to the school they ultimately transferred to. This means that I can create a proxy for how many transfer applications a given school received, but it is incomplete. Second, the open market transfer data do not always align with the payroll system data. Although 97% of the successful transfers in the open market transfer data are associated with transfers in the payroll data, only 80% of the transfers in the payroll data are associated with a transfer application, and only 63% with a successful transfer application. NYCDOE has not been able to identify the source of discrepancies or provide me with updated data, but has advised me that the administrative data from the payroll system is the more reliable source for actual teacher transfers. My primary results thus use the payroll system data (the payroll data also contain additional outcomes such as leaving NYCDOE entirely instead of transferring); the open market transfer data is used as suggestive supplementary evidence.

I also use data from surveys conducted with teachers by the NYCDOE as part of the accountability system. Surveys were conducted near the end of the schoolyear, after high-stakes tests were conducted but before the results were received. I use survey data from the 2006-07 through 2009-10 schoolyears. The schools asked questions on a broad range of issues, from contact with parents to school safety to interaction with principals.

I study schools in the first two years that accountability grades were released: the 2007-08 and 2008-09 schoolyears.[8]

3.2 Value-added estimation

To estimate teacher value-added, I created a matched-panel of student and teacher data.[9] I use the approach that has been experimentally validated in the economics of education literature (Kane and

[8] I do not use the data from later years of the accountability program for two main reasons: first, because changes were made to the program in the later years which made the thresholds less strict, including that outcomes began to depend not just on current grades but also on past performance; and, second, because data from those years are not included in the data files I was able to obtain from the NYCDOE.

[9] For the 2004-2005 through 2006-2007 school years, I matched teachers with classrooms based on a file maintained by the NYCDOE with student-level math and ELA teacher linkage data that has been verified by the schools. Based on guidance from the NYCDOE, for school years previous to 2004-2005, I matched elementary school students to teachers based on their homeroom identifiers, and middle school students to teachers based on course section identifiers.

Staiger, 2008). Appendix A describes the estimation in detail; I follow the recent literature and use Empirical Bayes shrinkage estimates (Jackson, 2012; Koedel et al., 2015). (Recent literature has highlighted the potential biases of the value-added approach; see Appendix A for discussion of why these biases should not be problematic here.) A primary strength of NYCDOE data is that the matched panel exists for eight years prior to the institution of accountability. This allows me to estimate value-added using data from the pre-accountability period and not conflate teacher quality with responses to accountability. As a result, value-added is only available for teachers who taught in tested grades before 2008.

Estimated value-added (VA) data is unfortunately only available for a subset of teachers (27% of all teachers in the sample, 21% of leavers, and 9% of joiners). I thus follow Jackson (2012) and calculate a predicted VA measure based on observable characteristics that is available for all teachers in the dataset. See Appendix A.1 for more details. Note that because I have a more limited set of observable characteristics than used in some of the existing literature (e.g., Jackson (2012)), the correlation between my predicted value-added measure and estimated value-added is relatively small (correlation of 0.06), although highly statistically significant. I thus focus on the estimated VA results as my main measure in the exposition, but use the predicted VA results to provide suggestive evidence of whether the results in the subsample with estimated VA data might extend more broadly.

3.3 Empirical Strategy

The RD approach adopted in this paper is similar to much of the literature studying the effects of accountability grades (e.g., Rouse et al. (2013), Chiang (2009)), and most closely follows Rockoff and Turner (2010), who use a similar specification to estimate how the NYCDOE accountability reforms affected short-run achievement. I estimate equations of the following form:

$$Y_{jt} = \alpha + \beta_g I^g_{jt} + \gamma h(S_{jt}) \times I^g_{jt} \times I^{type}_j \times I_t + \varepsilon_{jt} \qquad (1)$$

where j indexes teachers, t indexes time, g indexes accountability grades, Y_{jt} is the outcome variable of interest (e.g., an indicator that the teacher left the school), I^g_{jt} is an indicator for the grade received by a school, S_{jt} is the school's accountability score, $h()$ represents a flexible control function allowed to differ on either side of the grade threshold, and ε_{jt} is a mean 0 error term. I follow Rockoff and Turner (2010) in interacting the control function with an indicator for school type, I_j^{type}, as well as for the year, I_t, since the grade thresholds are all specific for school types and years.[10] My base specification for $h()$ follows Hahn et al. (2001) and much of the recent literature (e.g., Malamud and Pop-Eleches (2011)) in using a locally linear control function and a rectangular kernel.[11] I also explore robustness to parametric regression functions. All standard errors are clustered at the school level.

To increase statistical power, I follow Rockoff and Turner (2010)[12] and group the schools from the bottom thresholds together (C/D and D/F), and from the top thresholds together (A/B and B/C). The accountability pressures placed on schools (and the marginal increases in those pressures across grade thresholds) are much stronger at the bottom of the grade distribution than the top, and indeed, Rockoff and Turner (2010) find that accountability induced improvements in achievement at both bottom grade thresholds but not the top.[13]

Appendix B discusses selection of the base bandwidth used for the analysis. I also explore the sensitivity of the results to a range of bandwidths.

As with all RD analyses, the treatment effect under estimation is local to schools adjacent to the grade thresholds, and does not capture any universal effects of accountability. Since the analysis pools estimates from multiple cutoffs, the treatment effect represents the weighted average of the local effects at each individual cutoff, where cutoff values that are more likely to occur and that have more observations are given higher weight (Cattaneo et al., 2016). The identification assumption

[10] The qualitative findings are the same if I omit the interaction term, but the estimates are less precise.

[11] Cheng, Fan, and Marron (1997) show that the triangular kernel has boundary optimal properties, but, in practice, the results are not very sensitive to choice of kernel. Imbens and Lemieux (2008) and Lee and Lemieux (2010) recommend using a rectangular kernel and checking sensitivity to small bandwidths as an arguably more transparent method of putting more weight on observations close to the cutoff.

[12] Rockoff and Turner (2010) only adopt this specification when using locally linear control functions and small bandwidths, which is the specification adopted in this paper.

[13] Specifically, I assign all D (B) schools to the C/D or D/F (A/B or B/C) thresholds based on whether they were above or below the median score for other D (B) schools of the same year/schooltype. I then estimate equation (1), controlling for a linear trend in the accountability score for each schooltype-year combination, and including a dummy for whether a school received the lower grade at the grade threshold it was assigned to.

is that, conditional on the continuous metric underlying the grade, the grade itself is exogenous. Below I discuss the two primary potential violations of this assumption: manipulation of the running variable, and "*ex post* gaming" (selection into the analysis sample based on the grade received).

Density Evidence

A primary potential violation of the RD identification assumption is manipulation of the running variable. Manipulation is unlikely to be problematic in this setting because of the use of fixed grade thresholds and the fact that the underlying components of the score are all publicly verifiable and difficult to manipulate precisely (like test scores).[14] To investigate this further, I perform the density test suggested in McCrary (2008) and do not find evidence of bunching. Specifically, I use the number of schools within a 0.5-point or 0.1-point bandwidth as the dependent variable in equation 1 and the base bandwidth used in the analysis (5 score points), and none of the coefficients on the dummy for being on the lower-graded side of the grade threshold are statistically significant. See Appendix Table A1 for results.[15]

Sample construction and "*ex post* gaming"

My sample consists of all non-charter elementary, K-8, and middle schools that received accountability grades in the 2007-08 or 2008-09 school years (1,987 school-year observations).[16] I exclude the school-year observations where the school closed in the following year: this represents 5 school-year observations, all from the 2007-08 school year. As a result, unlike in most RD settings, one could be concerned about "ex post gaming" here: If administrators took accountability grades into

[14] The largest potential threat to identification is thus not in gaming of the score itself, but in gaming of the cutoffs by accountability officials. Although this concern is mitigated by the fact that the accountability program used round-number percentile cutoffs for thresholds and so would have had to manipulate the thresholds by large amounts to accommodate individual schools, it could still be the case that accountability officials changed the threshold to accommodate certain schools that they felt should receive given grades. However, since there are multiple schools receiving any given score, I rely on the fact that, even in the scenario that accountability officials changed thresholds to accommodate schools, that could only be for 1-2 schools, while the empirical results are driven by the many other schools at the threshold.

[15] Each column contains two cells with the results from separate regressions: the results at the top and bottom ends of the grade distribution. See Appendix Figure A2 for figures showing accountability scores on the X-axis and, on the Y-axis, the number of elementary schools. The figures are shown separately for each school type and year to show the full distribution, since the thresholds are specific to the school type and year.

[16] Charter schools did not receive accountability grades in 2007-08; in 2008-09, 40 charter schools received accountability grades, but are excluded because accountability may affect them differently and because I do not have other data for them.

account when selecting schools for closure, this would violate the identification assumption. However, most of the school closures took place far from the thresholds: only one of the schools that was closed during the sample period fell within a 5-point bandwidth of any grade threshold (the base bandwidth used in the paper), and there is no significant RD relationship between that closure and accountability grades (see App. Table A2, col. (1)). Since just one school-year observation is excluded, the results are also robust to bounding exercises assuming that outcomes at that school fall at the extreme ends of the distribution (see Section 6). Thus, *ex post* selection is not driving the results. I also exclude 6 school-year observations that do not appear in the teacher data (2 within a 5-point bandwidth of any grade threshold). These observations fall across the grade distribution, and there is no significant RD relationship between accountability grades and missing data (App. Table A2, col. (2)). The final sample has 1,976 school-year observations, 1,243 within 5 points of any grade threshold.

In addition to schools that were immediately closed, there were also schools that were phased out over time. According to Rockoff and Turner (2010), at the end of the 2007-08 school year, the NYCDOE announced that 7 schools would be either closed or phased out; the administrative data on school closures show that only 5 schools were closed immediately, leaving 2 schools that began phase out at the end of the first year of the program, and potentially more at the end of the second year. These phase-out schools are included in the base analysis sample. Although they do not represent a threat to identification, from an interpretation perspective, it is important to check whether different dynamics at these schools affect the results. Using a proxy for phase-outs,[17] I show that they do not affect the results: First, there is no significant RD relationship between phase-outs and accountability grades (App. Table A2, col. (3)); and, second, the results are robust to dropping phase-out schools (see Section 6).

My base analysis sample includes all schools. However, I also show my main results for a restricted sample that excludes the schools that had begun restructuring prior to the institution of

[17] Although the NYCDOE has administrative data on school closure dates, the NYCDOE does not track data on when school phase-outs began. However, I can proxy for the beginning of a phase-out by tagging schools that received accountability grades in one year but not in any subsequent year. This method tags 1% of schools, 75% of which closed in the ensuing 5-year period, implying that this proxy method seems reasonable. Results are also robust to using eventual closing as the proxy.

accountability (i.e., schools in year 2+ of restructuring in the 2007-08 school year). Restructuring often entails significant staffing shifts which are pre-determined to accountability and which could dampen their ability to respond to the school-based accountability system; thus, while the full sample has broader external validity, the restricted sample is interesting to look at for a more pure examination of how schools respond to one specific accountability system.[18]

3.4 Summary Statistics

Descriptive statistics for the base analysis sample (i.e., within 5 points of the grade thresholds) are presented in Table 1. (Statistics are very similar when excluding the restructuring schools, shown in Appendix Table A10). Panel A shows descriptive statistics about the sample of teachers teaching in the base analysis sample schools in the 2007-08 and 2008-09 school years. The two-year panel contains 50,616 unique teachers and 71,677 teacher-year observations. Roughly 27% of the teachers have math value-added data.[19] Baseline teacher value-added generally increases with the accountability grade.

Panel A of Table 1 shows that there is 11% teacher turnover across the sample period, with turnover increasing across accountability grades. Eight percent of the turnover is teacher retirements, 33% is transfers made between teaching positions in the NYCDOE, and 57% reflects departures from NYCDOE.[20] Turnover is generally higher among less-experienced and less-educated teachers.[21]

Table 1 shows that 15% of schools in the sample were part of the New York School Bonus Program (NYSBP), a pilot program of incentive pay for teachers started by the NYCDOE in the fall

[18] Six percent of all schools and 10% of schools within 6 points of a grade threshold had begun restructuring. Since the restructuring designation was pre-determined relative to accountability, it should not be related to accountability grades, and, reassuringly, col. (4) of App. Table A2 shows that there is no RD relationship between the two.

[19] Roughly 32% of them have either ELA or Math value-added. The reason that so many teachers do not have value-added data is that, in grades K-5, only grades 4-5 have usable value-added data (because only grades 3-5 are tested and one year of lagged test score is necessary for construction of value-added estimates), and in grades 6-8, only one of a student's approximately 5 teachers will be the subject teacher for math or for ELA; thus roughly 1/3 of teachers should be eligible to have ELA value-added data, and 1/3 for math.

[20] I cannot follow these teachers in the data: they could take teaching positions in other districts, take other non-teaching positions, stop working, or take non-teaching roles within the NYCDOE.

[21] See Appendix Table A11, which uses data from the pre-accountability era to understand the correlates of turnover. The negative correlation between experience and turnover could reflect the fact that there is no technical role for seniority in NYCDOE's open market system.

of 2007 in a small subset of schools. The program had limited impact, with no evidence of effects on overall teacher or student behavior (Fryer, 2013; Goodman and Turner, 2010; Springer, 2011). Since program participation was randomly assigned and determined both prior to and independently of the accountability system, it should be unrelated to accountability grades and should not affect the results; indeed, col. (5) of App. Table A2 shows that there is no RD relationship between grades and the program. I also include a NYSBP control in the vector of school-level controls in the regressions (described more below), but the results are invariant to exclusion of the control.

3.5 Balance Tests

Appendix Table A3 tests for balance in baseline characteristics by estimating equation 1 using baseline characteristics as the outcome variables. Each column contains two cells displaying the results from separate regressions, with the upper cell showing the results from the bottom end of the grade distribution (C/D and D/F thresholds pooled) and the lower cell showing the results from the top end (A/B and B/C pooled). The coefficient on the indicator for receiving the lower grade at the threshold is shown. I find roughly the number of significant coefficients that would be expected due to chance (5.8% of coefficients are significant at the 5% level). I thus follow Lee and Lemieux (2010) and perform a joint test of the significance of all coefficients, and reassuringly fail to reject the null that all of the coefficients are equal to 0 (the p-values are 0.30 for the top grade thresholds and 0.16 for the bottom grade thresholds). At the bottom of the grade distribution (the primary focus of the paper), the specific variables that are unbalanced are related to the racial composition of the school: the lower-graded schools have fewer black students and teachers, and more Hispanic students. These 3 variables are highly correlated with each other, with the correlation between percent black students and percent black teachers (percent Hispanic students) equal to 0.82 (-0.39). All regressions control for a vector of school-level variables that includes any variables unbalanced at the 5% level, in addition to other controls designed to improve precision.[22]

[22] School controls include: average student achievement from the previous year; the previous year's accountability score (second year only); the percent of students that are black, Hispanic, Asian, that receive free and reduced price lunch, and that are immigrants; the percent of teachers that are black, Hispanic, or Asian; fixed effects for school size; five-year average school turnover before the institution of the accountability system; and whether the school was part of the NY school bonus program. For the turnover regressions, I also include a vector teacher-level covariates that

4 Results

4.1 Turnover Results

I begin with graphical evidence. The left column of Figure 1 plots average residual turnover against accountability scores. The top graph shows the results at the bottom of the grade distribution (C/D and D/F pooled), the lower graph shows the top pooled (A/B and B/C). To create residual turnover, I regress an indicator for whether a teacher left a school on a vector of covariates.[23] I then group schools according to their accountability scores relative to the grade threshold, with each dot representing the schools within a 1-point bandwidth, and plot the average residual turnover at the end of the school year on the average accountability score received at the beginning of the year.[24] For ease of interpretation, I also plot a locally linear regression line in the figures, fitted to the raw teacher-level data and fitted separately on either side of the grade threshold. Shaded areas represent 95% confidence intervals around the line.

At the bottom of the grade distribution (C/D and D/F), lower-graded schools appear to have locally lower turnover. In contrast, at the top of the grade distribution, there is virtually no discontinuity at the threshold.

To test for the significance of these results, columns (1) and (2) of Table 2 present the regression results, calculated from estimation of equation (1) using an indicator for whether a teacher left the school at the end of the school year as the dependent variable. Each cell contains the results from a separate regression, with the results from the bottom end of the grade distribution shown first. All regressions control for the vector of controls outlined in Section 3.5. Column (1) shows the full-sample results, whereas column (2) shows the results in the non-restructuring sample.

Consistent with the graphical evidence, the regressions show that at the bottom end of the grade distribution lower accountability grades decrease turnover by roughly 3 percentage points, signif-

the literature has shown to influence teacher turnover, including fixed effects for teacher experience and age, teacher education level, teacher race, and teacher gender.

[23] The covariates used for creating the residuals are the same used in the regressions, described below.

[24] Note that, because density varies along the x-axis, the number of schools in each dot varies; Appendix Fig. A3 shows versions of the graphs where the number of schools per dot is fixed instead of the bandwidth per dot (so each dot represents the average score and the average turnover across 10 schools).

icant at the 5% level. This reduction in turnover should provide a direct benefit to low-graded schools, since turnover has been shown to decrease student achievement (Ronfeldt et al., 2013). The magnitude of the effect is economically meaningful, equal to roughly 20% of average turnover amongst those schools. To put the magnitude in the context of the literature, studies on the elasticity of turnover to wage changes find elasticities ranging from roughly -1 to -1.5 (Hanushek et al., 2004; Hendricks, 2014; Dolton and von der Klaauw, 1995) to roughly -3 to -4 for women and -3 to -8 for men (Imazeki, 2005; Clotfelter et al., 2008), implying that the 20% decrease in turnover seen here would be equivalent to an increase in salary of anywhere from 2.5-7% to 13-20%. The magnitude of the point estimates is substantial, but seems reasonable in the context of the literature on mobility responses to other reforms, including accountability, for which estimates of the magnitude of the effects on teacher turnover are in the range from 25% to 65%.[25] While the point estimates are large in magnitude, since turnover is a rare event, the confidence intervals are relatively wide.

In contrast, at the top end of the grade distribution, where accountability pressures have less bite, grades do not affect turnover: the estimate is small in magnitude (0.5 percentage points, col. (1)) and not statistically distinguishable from 0.

I examine the robustness of the turnover results in Section 6.1.

4.1.1 Turnover Placebo Test

The credibility of the RD design rests on the assumption that schools are as if randomly assigned at the grade thresholds. To examine the validity of this assumption, I also perform a placebo test, checking whether there are any baseline differences in turnover between schools on either side of grade thresholds. Thus, I estimate the exact same regression model, but using turnover from the *previous* year, i.e., from the year before the accountability grade was received, as the dependent

[25] Most of the other literature on school accountability finds effects in the other direction, where the differences in the sign of the effects likely comes from reasons that will be discussed in Section 7, but the magnitudes are substantial: Feng et al. (2010) estimate that being shocked to receive an F grade in Florida increases turnover by 42%, Clotfelter et al. (2004) estimates that being labeled a low-performing school increases turnover by 25%, and Gjefsen and Gunnes (2016) estimate that the institution of school accountability increased teacher turnover by roughly 45-65%. One can also benchmark against other types of policies and reforms and again the estimates do not seem unreasonable. For example, Smith and Ingersoll (2004) find that giving new teachers a mentor in their field or encouraging collaboration decreases turnover of first-year teachers by 30% and 43% respectively.

variable.[26] Column (3) of Table 2 shows the placebo tests using both years of data, while column (4) shows the placebo from the first year of the program only in order to limit to pre-accountability data.[27] Reassuringly, none of the placebo coefficients are statistically significant, and the magnitudes are relatively small. The right column in Figure 1 presents the placebo results graphically, with consistent results.

4.2 Teacher Sorting Results

I now look at how accountability grades affect the quality of the teachers who leave schools (leavers) and the teachers who join in the next year (joiners). Note that the results in this section should be interpreted as suggestive because the samples of teachers are small, especially the sample of joiners with estimated value-added data. The placebo tests presented in the next subsection are also not dispositive.

Panel A of Table 3 presents the RD estimates of the effect of receiving a lower grade on the average quality of leavers and joiners, calculated by estimating equation (1) using teacher value-added as the dependent variable and using either the leavers or the joiners as the sample. I use mathematics value-added as my value-added measure since the literature has shown that teacher fixed effects in mathematics tend to have more predictive power over student outcomes than ELA fixed effects (e.g., Jacob and Lefgren (2008), Jackson and Bruegmann (2009)). This probably reflects the fact that, while most students learn language skills from many sources (e.g., their parents, the television), the primary source of math knowledge for many students is their teachers (Gates Foundation, 2010).[28] Figures 2 and 3 depict the estimated VA results graphically, with consistent results.

[26] Since regressions are run at the teacher level, using previous-year school turnover as the dependent variable means using teachers who were teaching in the school in the previous year as the sample. Note that the vector of control variables are taken from the previous year as well so that the control variables are predetermined, not measured after the (placebo) outcome.

[27] Using both years has the advantages of testing for baseline differences across the full sample that identifies the actual results, and of having the same power for rejecting the null as the actual results. However, some of the previous papers in the accountability literature use data from the first year of the accountability system only so that the falsification exercise is performed purely with pre-accountability data, so I perform that test as well. Placebo results for the second year only are the same as well. Placebo results are shown using the full sample, but the results in the non-restructuring sample are nearly quantitatively identical.

[28] The results using ELA value-added are statistically weaker and less robust, and available upon request. Note that roughly 70 percent of teachers with any value-added data have both math and ELA value-added.

There are two ways to examine leaver quality: look at how leaver quality varies on either side of the grade threshold, or estimate the turnover regressions from the previous section separately for teachers with different value-added. The former is shown in Table 3, while the latter approach is shown in Table 4.[29] I focus on the former specification for the exposition since it enables an easier comparison with the joiner quality results, but the results are consistent.

Beginning with the bottom end of the grade distribution, where we saw that lower accountability grades decreased average turnover, there is not strong evidence that turnover varies by quality (columns (1) through (4)): the point estimates are negative using estimated VA (implying that higher VA teachers were *more* likely to stay), but positive using predicted VA. None of the coefficients using this specification are statistically significant, with the estimated VA result significant at the 10% level in the Table 4 specification. However, there is suggestive evidence of an impact on joiner quality. The point estimates in columns (5) through (6) of Table 3 suggest that lower-graded schools attract joiners with 0.9 std. dev. higher estimated value-added, statistically significant at the 5% level. Since estimated VA is only available for a small sub-sample of the joiners, it is reassuring that the results using predicted VA, which is available for the full sample, are qualitatively consistent, if weaker, potentially reflecting the low power of the predicted VA measure in predicting VA (cols. (7) - (8)). It is also reassuring that there is no significant selection into having estimated VA data by accountability grade (Panel B, column (4)). Overall the results provide suggestive evidence of positive sorting implications for lower-graded schools, as lower-graded schools may attract higher-quality joiners relative to leavers.

In contrast, at the top end of the grade distribution (i.e., A/B and B/C), there is no evidence of a positive effect on quality, and some evidence that lower accountability grades may, if anything, decrease the quality of joiners relative to leavers. However, the results are only suggestive, as the leaver result only shows up in the predicted VA data and the joiner only in the estimated VA.[30]

[29] Specifically, the table presents results from estimating equation (1) where the RD control function and the indicator for receiving the lower grade are both interacted with a teacher characteristic.

[30] Note that, unlike Table 3, Table 4 does not seem to suggest that teachers with higher predicted VA are more likely to leave. The discrepancy reflects the functional form, i.e., the fact that I need to convert to a binary variable for the Table 4 specification and so use "above-median" for the split; if I split at other places in the distribution (e.g., 25th percentile, 90th percentile) then, consistent with Table 3, teachers with higher predicted VA do appear to have higher turnover,

Table 4 and Panel B of Table 3 also show results for other teacher characteristics (experience and education); none of the results are statistically significant. In general, no single characteristic proxies well for value-added (Rivkin et al., 2005; Hanushek et al., 2010), and so I do not see these results as inconsistent with the others.[31]

4.2.1 Placebo Tests

Appendix Table A4 presents placebo results regressing the average *previous-year* characteristics of leavers and joiners on accountability grades, shown for both years of data (panel A) and the first year only (Panel B). There are roughly the number of significant coefficients that would be expected due to chance in Panel A (none significant at the 5% level and 2/16 significant at the 10% level), and a somewhat higher percentage for Panel B (2/16 significant at the 5% level). See the right column of Figures 2 and 3 for placebo graphs. The robustness of the joiner results is examined in Section 6.2.

4.3 Summary of Results

At the bottom end of the grade distribution, receiving a lower accountability grade causes teacher turnover to fall, and improves the quality of the joiners hired relative to the leavers. These results thus provide a more hopeful story for accountability than suggested by policy makers, implying that, through their labor market effects, accountability systems can in some cases benefit, not harm, the most disadvantaged schools. In contrast, at the top end of the grade distribution, the teacher labor market effects of receiving a lower grade are not positive, and, if anything, are negative. In the next section, I examine which mechanisms could explain these findings.

5 Mechanisms

The finding that receiving a lower accountability grade causes teacher turnover to decrease is somewhat unexpected. Turnover decisions can in general either reflect supply-side factors, demand-side

although the results are never significant.

[31] For experience, an indicator for having at least 4 years of experience is shown since the quality gains to experience generally taper after 4 years (Boyd et al., 2008; Kane et al., 2008; Rivkin et al., 2005; Rockoff, 2004), and, in the NYCDOE data, there is no significant relationship between experience and value-added after four years, but the results look similar with other experience measures. The correlation between teacher value-added and masters degrees is not statistically significant in the NYCDOE data, and is generally tenuous, sometimes even negative (Rivkin et al., 2005).

factors, or some combination of the two. Here, supply-side factors would mean that the desirability of working at these schools has increased, which I will refer to as the *job desirability* hypothesis. Demand-side factors would mean that external employers do not want to hire teachers from lower-graded schools, which I will refer to as the *stigma* hypothesis.

The *job desirability* hypothesis is somewhat counterintuitive, since many of the factors one might first associate with lower grades (such as lower prestige, higher pressure, higher risk of future closure, etc.) are negative. Thus, the *job desirability* hypothesis requires that between the beginning of the year (when schools receive grades) and the end of the year (when teachers make turnover decisions), lower-graded schools respond to accountability pressures by making changes that make the school a more desirable place to work. One potential channel would be performance improvements: the previous literature has shown that schools respond to accountability pressures by improving their performance (Carnoy and Loeb, 2002; Hanushek and Raymond, 2005; Chiang, 2009; Rockoff and Turner, 2010). Accountability pressures could also motivate principals to do things that are attractive to teachers, such as focusing more on teacher development, offering more opportunities for teachers to collaborate, or providing more autonomy. Relatedly, accountability pressures could cause principals to focus more on hiring and retaining the best teachers (Shipps and White, 2009). Alternatively, lower grades could foster collaboration as teachers work together to improve. Teachers could even like the challenge of the lower grade, perhaps because it makes them feel that they are making a difference for students.[32]

The evidence presented so far seems more aligned with the *job desirability* hypothesis, although is far from definitive. The fact that lower-graded schools attract higher-quality joiners provides some evidence for the “job desirability” hypothesis by suggesting that the lower-graded schools or principals made some positive changes or increased their recruitment efforts. The fact that the decrease in turnover is only seen at the bottom of the grade distribution may also support the *job desirability* hypothesis, as the change in accountability pressures when crossing a grade threshold

[32] Indeed, in a related context (subgroup-specific accountability policies), Shirrell (2016) finds that, when a school is held accountable for the performance of black students, black teachers’ turnover falls, with one potential mechanism being that the teachers’ motivation increases when they see that the school is trying to improve the performance of black students.

were larger at the bottom end of the grade distribution, and thus any pressure-induced increases in job desirability should have been larger there too. In contrast, although possible, it is not clear why one would expect *ex ante* that stigma would change more across thresholds at the bottom end of the grade distribution than at the top.

I now provide further evidence on mechanisms. I first provide evidence for the plausibility and potential channels for the "job desirability" hypothesis. I cannot distinguish between all of the potential channels outlined above since I do not have data on them all and they all are likely highly correlated, but I do provide evidence that supports the hypothesis that there are some ways that the lower-graded schools were improving, and suggestive evidence that these changes are attractive to teachers. I then use data on transfers and transfer applications to provide additional tests of the *stigma* hypothesis. I do not find evidence supporting the *stigma* hypothesis.

5.1 Mechanisms for the supply-side (job desirability) hypothesis

Rockoff and Turner (2010) show that accountability pressures spurred achievement improvements at lower-graded schools at the bottom of the grade distribution in the NYCDOE even within the same year, so that is a plausible mechanism for improved job desirability. Results of teacher surveys conducted by the NYCDOE at the end of the school year can provide further evidence on mechanisms. The NYCDOE grouped questions in the survey by topical area (e.g., principal leadership, campus environment); I created indexes for each topical area equal to the average standardized responses across all the questions in each section.[33] Panel A of Table 5 presents the results. At the lower grade thresholds, the strongest impacts are on the principal leadership index, which is roughly 0.3 sd more positive at lower-graded schools than higher-graded schools, although only significant at the 10% level (col. (1)). Panel B shows the detailed questions that make up the principal leadership index. Teachers were more likely to think that their principals supported them, and that their principals were effective managers who made the school run smoothly. Since the differences in perceived principal leadership were not present at baseline, this suggests that low grades encouraged principals

[33] Sometimes two sections covered the same topic and so I grouped those sections together. All responses were normalized so that more positive was better.

to work harder at their jobs and that teachers appreciated these changes.[34]

We do not see similar results at the top end of the grade distribution: accountability pressures may not have been large enough to induce principal responses, perhaps explaining why there also were no turnover impacts at those thresholds. In fact, if anything, there is suggestive evidence that the effect on survey responses were negative, consistent with the general trend that, unlike at the bottom of the grade distribution, at the top, low accountability grades seem to have negatively impacted schools, presumably because the accountability pressures were not strong enough to spur positive changes by principals and teachers.

Appendix Table A5 tests for other changes at lower-graded schools, testing for changes to the number of joiners, staff size, enrollment, class size, and principal turnover. (The change in number of leavers, i.e., turnover, is included as column (1) to compare with the joiner effects.) There are no significant impacts on any of these variables.[35] Of course, other changes could be happening for which I do not have data.[36]

In order for changes such as higher student performance or better principal leadership to increase job desirability, teachers must value them. Table 6 provides suggestive evidence that teachers do value these types of changes by showing that, in general, when schools have higher performance or principal leadership survey scores (conditional on previous-year performance or principal leadership

[34] Across the 10 variables shown, no placebo results are significant at the 5% level, and one (the course offerings index in panel A) is significant at the 10% level. Note that one potential concern with these results is that teacher survey results are an input into future accountability grades. Although they only represent roughly 3% of grades (teacher, student, and parent surveys together represent 2/3 of the school environment score, which represents 15% of the overall score), some teachers may still have answered questions strategically to try to affect their future grades, and the likelihood of answering strategically could vary with accountability grades. Since all survey questions affect the accountability grade equally, this could cause a bias in the average responses across all questions, but is less likely to cause more positive responses to some questions than others. Thus, the fact that there are much more positive responses for principal leadership than for, say, safety, likely reflects true effects, not strategic responses. The fact that there are some negative coefficients is also suggestive that teachers were not responding strategically.

[35] The number of joiners falls by 1% at lower-graded schools, which is not enough to offset the fall in leavers, and thus the number of teachers increases by 2% (not statistically significant). Enrollment increases by 1% and class size falls by 1%. There is no effect on principal turnover, and the turnover results are also robust to excluding principals who left [coefficient of -3.2% at the bottom of the grade distribution when leaving principals are excluded]. All regressions are weighted by staff size.

[36] One thing that could change is teacher expectations of future school closures. When schools close in the NYCDOE, teachers are not fired. If they cannot find permanent positions, they are given work as substitute teachers. If teachers prefer substitute work, they could stay at lower-graded schools hoping that the schools will be closed in the future. I do not view this hypothesis as very plausible because (1) all of the closures had already been announced before teachers made turnover decisions, and so they would have needed to anticipate closures a full year in the future, and (2) anecdotal evidence suggests that teachers dislike being substitutes.

scores), those schools also experience lower teacher turnover.[37]

Further evidence on mechanisms can be provided by looking at whether the RD effects are larger at schools that we would predict would respond more positively to accountability pressures. The teacher survey results suggest an important role for principals; high-quality principals may be able to better channel accountability pressures into positive changes for their schools. Consistent with this hypothesis, Table 7 shows that the decreases in turnover are 6 percentage points larger among schools with above-median baseline principal leadership (as measured in the teacher surveys). This difference is substantial in magnitude, and significant at the 5% level. This suggests that schools with good leadership may have been better able to translate accountability pressures into positive changes that the teachers appreciate. The results for heterogeneity in the effects on survey responses about principal leadership and joiner quality are not statistically significant, although the point estimate for the survey responses is large, suggesting three times as large an effect for schools with better principals at baseline. For joiner quality, the point estimate is actually negative, but confidence intervals are very wide and so I cannot reject that the improvements in joiner quality were substantially larger at schools with higher-quality principals.

5.2 Additional Tests: Destinations and Transfer Applications

The *stigma* and *job desirability* hypotheses have different implications for how the results will vary by teachers' destinations. Since external (non-NYCDOE) employers are unlikely to look up school accountability grades, stigma should primarily affect intra-district transfers, whereas the job desirability hypothesis could affect all types of turnover. Table 8 shows the turnover results by destination, with column (1) replicating the overall result from Table 2, and columns (2) through (4) showing the results separately for retirement, transfers between NYCDOE schools, or leaving the NYCDOE.[38] The turnover result is driven almost entirely by fewer teachers leaving NYCDOE (col.

[37] Note that the regressions pool across all schools and teachers in my sample, not just those at the bottom or top of the grade distribution, but the results are different if one limits to schools in one area of the grade distribution. Principal leadership results are also similar if one does not condition on previous-year leadership. The correlations with joiner VA are more mixed than the turnover correlations, but joiner VA is also much less precise.

[38] Teachers who leave the NYCDOE could be changing professions, taking a short stint away from teaching, transferring to a different district, or taking non-teaching roles within the NYCDOE.

(2)), which accounts for over 75% of the decrease in turnover at lower-graded schools, somewhat larger than the share of turnover driven by departures from the NYCDOE (50%). In contrast, within-district transfers (col. (4)) represent a smaller percentage of the effect than of overall turnover. (Note that in neither case can we reject equality.)

Columns (5) through (8) bring in transfer application data from the open market transfer system to shed further light on the mechanisms (note that I consider these analyses suggestive due to the caveats with the open market data raised in Section 3.1). Under the stigma hypothesis, transfer applicants from lower-graded schools should be less likely to receive job offers and thus to ultimately transfer. The number of transfer applications received by lower-graded schools should likely fall as well. This is not what the data suggest: using either the open market transfer data (col. (6)) or the payroll data (col. (7)) to define which applicants ultimately transferred, lower grades do not decrease the transfer rate among applicants, and lower-graded schools do not receive any fewer applications to transfer, with both point estimates in fact positive (i.e., running against the stigma hypothesis).

Note that at the top of the grade distribution, consistent with the general trend that, if anything, lower accountability grades seem to make schools worse off through the labor market impacts, we see more transfer applications submitted (significant at the 10% level), although this is hard to interpret due to the discrepancies between the open market transfer and payroll data; in the payroll data, transfers themselves remain flat.

6 Robustness of the RD Results

6.1 Robustness of the Turnover Results

Table 9 shows robustness of the turnover results to the RD specification used. Column (1) shows the results estimated without including baseline covariates; the coefficient remains negative, with the magnitude still large but smaller than the base specification (shown in column (2)), and standard errors much larger. Columns (3) through (6) show the results using linear specifications with a range of bandwidths (specifically, 50% and 200% of the base bandwidth, and the base bandwidth +/-1). Columns (7) and (8) show that the results are qualitatively similar if one uses a parametric regression function (either quadratic or cubic in the accountability score, estimated separately by grade using a bandwidth of 200% the base bandwidth). Column (9) shows the specification linearly for all of the components of the accountability score separately instead of the composite score. The coefficient is smaller in magnitude but maintains its sign and I cannot reject equality with my base specification.[39] Column (10) shows the results after collapsing the data at the school level, weighted by the number of teachers at the school.

Given the noise in the graphs, one might be concerned that there are random breaks in the regression function. Per Lee and Lemieux (2010), I perform a specification test, testing for discontinuities at points other than the grade thresholds, and present p-values in Table 9.[40] Reassuringly, the test statistic is not rejected in any of the specifications.

Appendix Table A6 shows that the results are not driven by different dynamics at schools that were being phased out, as the results are robust to excluding those schools (col. (2)), nor are the results driven by sample selection due to the one school within five points of a grade threshold that was closed and thus excluded from the analysis, since the results are robust to a bounding exercise where we include the closing school and assign it turnover at the extremes of the distribution (the 99th percentile or 1st percentile, cols. (3)-(4)). The results are also robust to counting midyear departures as turnover (col. (5)).

Given the density and placebo tests presented earlier, I do not think that gaming is driving the results. However, looking at the results for 2007-08 and 2008-09 separately can also provide more insight: 2007-08 was the first year of the accountability system, and so it is especially unlikely that schools could have manipulated their scores around the cutoffs in that year.[41] Columns (6)-(7) show that, reassuringly, the results are similar when one estimates equation 1 separately for the 2007-08

[39] My base specification adopts the more standard approach in the literature of using the running variable itself as opposed to its underlying components.

[40] Specifically, I test for whether the discontinuities at all 1-point intervals from the grade threshold are all equal to zero. Results are robust to different interval widths.

[41] See Rockoff and Turner (2010) for a complete timeline of events. It is unlikely that schools knew what their 2007 accountability grades would be in advance. In April 2007, the NYCDOE informed principals of the progress report methodology and gave principals pilot progress reports based on 2005 and 2006 results. These reports did not contain letter grades, only numeric scores, and did not inform principals about how the numeric scores would be mapped to grades. The pilot reports also omitted other key information (e.g., peer groups, environmental scores) that would ultimately affect the schools score. Anecdotal newspaper evidence indicates that some principals were surprised to receive low grades.

and 2008-09 school years. Appendix Table A7 also shows that the estimates are robust to excluding schools that fall directly on a grade threshold.[42]

One may also wonder if the results vary at the different thresholds that are grouped in the analysis (e.g., C/D vs. D/F). Columns (1) and (2) of Appendix Table A8 shows that the main turnover results are qualitatively consistent across thresholds. At the bottom end of the grade distribution, the magnitude of the coefficient is larger at the D/F threshold than the C/D, but I cannot reject equality.

6.2 Robustness of the Joiner Quality Results

Table 10 repeats the analyses from Table 9 to examine the robustness of the joiner quality results. The results at both the top and bottom ends of the grade distribution are relatively robust across specifications, although the coefficient magnitude and significance vary somewhat. The positive result at the bottom end of the grade distribution is not robust to the cubic specification. However, the parametric specifications (quadratic and cubic) may allow too much flexibility given the noise in the data, as suggested by the fact that the p-value for the specification test is rejected in the quadratic specification (this test can also be seen as a test for whether the regression function is well approximated by the control function within the bandwidth). Appendix Table A9 shows that the results are robust to including the schools undergoing restructuring or excluding the schools that were being phased out, and that the results were qualitatively similar in 2008 and 2009.[43] One potential concern with the results would be if they were driven primarily by teachers coming from the schools that were closed for having low accountability grades. Since there were no restrictions on where these teachers could be hired, this would not be an internal validity concern, but could be an external validity concern if there were more teachers from closed schools in the NYCDOE than there typically are in other settings. However, teachers from closing schools do not seem to drive the results: these teachers represent less than 2% of the overall population of joiners and less than

[42] Fort et al. (2016) outline potential bias that can arise in multi-threshold RD settings when all thresholds have exactly one observation located precisely on the threshold (i.e., exactly one observation for which the running variable takes the value 0). Here, that concern likely does not apply since only 17% of the thresholds I analyze (4/24) have exactly one school at the threshold, but, to be conservative, Appendix Table A7 shows that the estimates are nearly identical when estimated excluding the observations at the cutoff – which is the recommended strategy to address the potential bias.

[43] The 2009 results are shown without controls only since the versions with controls are not identified due to small sample sizes and the large number of controls.

5% of the joiners with value-added data, and the results are robust to omitting these teachers from the analysis (col. (5)).[44]

7 Discussion

In this section, I discuss my findings in the context of the literature. This paper suggests that accountability pressures help low-performing schools by decreasing turnover and potentially improving teacher quality. The findings are inconsistent with the majority of the previous literature, which has largely found that accountability pressures hurt low-performing schools by accelerating turnover (Feng et al., 2010; Clotfelter et al., 2004; Gjefsen and Gunnes, 2016), more so at the bottom of the grade distribution in Feng et al. (2010).[45]

What might explain the differences between my results and some of the other findings in the literature? Part of the explanation is likely the stakes: Gjefsen and Gunnes (2016) study an accountability regime that does not have high-stakes rewards and sanctions, which likely explains the differences, as their results are more analogous to my results at the top end of the grade distribution than the bottom. However, other papers, such as Feng et al. (2010), also study high-stakes settings and find different results. My results on heterogeneity by principal leadership capacity suggests that variation in principal capacity and control across settings may contribute to the difference in results. For example, in the Florida accountability system studied by Feng et al. (2010), administrators do not seem to directly close low-graded schools as a disciplining mechanism like they do in the NYCDOE (Chiang, 2009; Feng et al., 2010). So, it may be the case that the Florida system does not place as much pressure on principals as the NYCDOE system. However, this is speculative, as I do not have enough information on the institutional differences regarding principals to say for sure, and

[44] Columns (3) - (8) of Appendix Table A8 show the main joiner and leaver results separately by grade threshold. The results are qualitatively consistent across grade thresholds, although the results, especially at the lower end of the grade distribution, should be treated as suggestive at best due to the small samples involved. Note that, when using estimated VA as the outcome variable (columns (7)-(8)), the D/F threshold sample has a particularly small sample. The result is shown for transparency, but the coefficient should be interpreted with great caution due to the small sample and potential for overfitting. Indeed, the result is sensitive to the covariates included: if we limit the school covariate vector to only the covariates that were unbalanced in the balance tests, the coefficient falls from 2.1 to 1.7 (p-value of .06), and when we remove all school covariates it falls to 0.83 (p-value of 0.20).

[45] In a related setting, Boyd et al. (2008) exploit within-school variation to find that, when New York state introduced high-stakes testing for fourth grade teachers, turnover among fourth grade teachers fell.

there are of course many other differences between the settings.[46] Thus, I cannot definitively say why the results differ; with just a few settings, the problem is not identified.

8 Conclusion

In this paper, I present evidence that accountability pressures impact the teacher labor market. At the bottom end of the grade distribution (the C/D and D/F thresholds), I find that accountability positively impacts lower-graded schools by decreasing turnover and by increasing the quality of joiners. These results echo the results from the earlier literature showing that lower accountability grades also improve academic performance (e.g., Rockoff and Turner (2010)). A plausible explanation here is that teachers actively choose to stay in the lower-graded schools because job desirability increases at those schools, perhaps because principals – especially high-quality principals – put more effort into leading their schools and into transforming the accountability pressures into positive change. In contrast, at the top end of the grade distribution (the A/B and B/C thresholds), where the accountability pressures are more mild, there is no evidence of positive impacts, and instead evidence that, if anything, receiving a lower accountability grade hurts schools through the labor market impacts, with suggestive evidence of a decrease in the quality of joiners relative to leavers.

These results raise an important question: When will accountability have positive impacts through its labor market effects, and when negative? The results suggest two ingredients for when accountability might lead to positive effects: in cases where (a) accountability is high-stakes enough to motivate schools to change, and (b) principals are good leaders and have the latitude to implement changes. These hypotheses are motivated by the fact that (a) I only see positive impacts at the bottom end of the grade distribution, where the stakes were higher, and (b) the positive turnover effects at the bottom end of the grade distribution are primarily driven by schools led by high-quality principals. Of course, these hypotheses are still speculative, and the question remains why other papers

[46] My findings also contrast with those of Clotfelter et al. (2004), who use a difference-in-differences approach to estimate the effect of the institution of accountability in North Carolina and find that accountability accelerated teacher turnover at low-performing schools. Again, institutional differences could have played a role; for example, North Carolina linked *teacher-level* incentives with school accountability ratings, whereas the NYCDOE system only used school- and principal-level incentives. It is also possible that their results are partially explained by other reforms instituted concurrently with accountability, such as streamlining the process of teacher dismissals and dramatically changing salary structures and tenure requirements.

in the literature looking at high-stakes systems find dissimilar results. I cannot definitively say why the results differ, primarily because, with just a few settings, the problem is not identified. One area for further research is to investigate the reasons for the differences, and in particular, the extent to which they reflect the context and design features of the accountability system. A more thorough understanding of these features would enable policymakers to design accountability systems that better improve the performance of disadvantaged schools.

References

Boyd, D., H. Lankford, S. Loeb, J. Rockoff, and J. Wyckoff (2008). The narrowing gap in New York City teacher qualifications and its implications for student achievement in high-poverty schools. *Journal of Policy Analysis and Management 27*(4), 793–818.

Boyd, D., H. Lankford, S. Loeb, and J. Wyckoff (2008). The impact of assessment and accountability on teacher recruitment and retention: Are there unintended consequences? *Education Finance and Policy 1*(36).

Carnoy, M. and S. Loeb (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evalation and Policy Analysis 24*(4), 305–331.

Cattaneo, M. D., R. Titiunik, G. Vazquez-Bare, and L. Keele (2016). Interpreting regression discontinuity designs with multiple cutoffs. *The Journal of Politics 78*(4), 1229–1248.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014). Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American Economic Review 104*(9), 2633–2679.

Chiang, H. (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics 93*(9), 1045–1057.

Clotfelter, C., E. Glennie, H. Ladd, and J. Vigdor (2008). Would higher salaries keep teachers in high-poverty schools? evidence from a policy intervention in north carolina. *Journal of Public Economics 92*(5), 1352–1370.

Clotfelter, C., H. F. Ladd, J. L. Vigdor, and R. A. Diaz (2004). Do school accountability systems make it more difficult for low-performing schools to attract and retain high-quality teachers? *Journal of Policy Analysis and Management 23*(2), 251–271.

Dolton, P. and W. von der Klaauw (1995). Leaving teaching in the uk: A duration analysis. *The Economic Journal*, 431–444.

Falch, T. and M. Rønning (2007). The influence of student achievement on teacher turnover. *Education Economics 15*(2), 177–202.

Feng, L., D. N. Figlio, and T. Sass (2010). School accountability and teacher mobility. Technical report.

Figlio, D. N. and C. E. Rouse (2006). Do accountability and voucher threats improve low-performing schools? *Journal of Public Economics 90*(1), 239–255.

Fort, M., A. Ichino, and G. Zanella (2016). On the perils of stacking thresholds in rd designs.

Fryer, R. G. (2013). Teacher incentives and student achievement: Evidence from new york city public schools. *Journal of Labor Economics 31*(2), 373–407.

Gjefsen, H. M. and T. Gunnes (2016). The effects of school accountability on teacher mobility and teacher sorting.

Goodman, S. and L. Turner (2010). Teacher incentive pay and educational outcomes: Evidence from the nyc bonus program. program on education policy and governance working papers series. pepg 10-07. *Program on Education Policy and Governance, Harvard University*.

Hahn, J., P. Todd, and W. Van der Klaauw (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica 69*(1), 201–209.

Hanushek, E., S. Rivkin, D. Figlio, and B. Jacob (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review 100*(2), 267–271.

Hanushek, E. A., J. F. Kain, and S. G. Rivkin (2004). Why public schools lose teachers. *Journal of Human Resources 39*(2), 326–354.

Hanushek, E. A. and M. E. Raymond (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management 24*, 297–327.

Hendricks, M. D. (2014). Does it pay to pay teachers more? evidence from texas. *Journal of Public Economics 109*, 50–63.

Hensvik, L. (2012). Competition, wages and teacher sorting: Lessons learned from a voucher reform. *The Economic Journal 122*(561), 799–824.

Imazeki, J. (2005). Teacher salaries and teacher attrition. *Economics of Education Review 24*(4), 431–449.

Imbens, G. and K. Kalyanaraman (2012). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies 79*(3), 933–959.

Imbens, G. and T. Lemieux (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics 142*(2), 611–650.

Jackson, C. K. (2009). Student demographics, teacher sorting, and teacher quality: Evidence from the end of school desegregation. *Journal of Labor Economics 27*(2), 213–256.

Jackson, C. K. (2012). School competition and teacher labor markets: Evidence from charter school entry in north carolina. *Journal of Public Economics 96*(5), 431–448.

Jackson, C. K. and E. Bruegmann (2009). Teaching students and teaching each other: The importance of peer learning for teachers. *American Economic Journal: Applied Economics 1*(4), 85–108.

Jacob, B. and L. Lefgren (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics 26*(1), 101–136.

Jones, M. G., B. D. Jones, B. Hardin, L. Chapman, T. Yarbrough, and M. Davis (1999). The impact of high-stakes testing on teachers and students in North Carolina. *The Phi Delta Kappa 81*(3), 199–203.

Kane, T. J., J. E. Rockoff, and D. Staiger (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review 27*, 615–631.

Kane, T. J. and D. O. Staiger (2008). Estimating teacher impacts on student achievement: An experimental evaluation. Nber working paper 14607.

Kirtley, K. (2012). *High stakes testing in lower-performing high schools: Mathematics teachers' perceptions of burnout and retention*. Ph. D. thesis, University of Colorado.

Koedel, C., K. Mihaly, and J. E. Rockoff (2015). Value-added modeling: A review. *Economics of Education Review 47*, 180–195.

Lee, D. and T. Lemieux (2010, June). Regression discontinuity designs in economics. *Journal of Economic Literature 48*, 281–355.

Li, D. (2011). School accountability and principal mobility: How no child left behind affects the allocation of school leaders. Working Paper, MIT.

Ludwig, J. and D. L. Miller (2007). Does Head Start improve children's life chances? Evidence from a regression discontinuity design. *Quarterly Journal of Economics 122*(1), 159–208.

Malamud, O. and C. Pop-Eleches (2011). Home computer use and the development of human capital. *Quarterly Journal of Economics 126*(2), 987–1027.

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

Peterson, P. E. (2006). The A+ plan. *Reforming Education in Florida: A Study Prepared by the Koret Task Force on K–12 Education. Stanford, Calif.: Hoover Institution*.

Rivkin, S., E. Hanushek, and J. Kain (2005). Teachers, schools, and academic achievement. *Econometrica 73*(2), 417–458.

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

Rockoff, J. and L. J. Turner (2010). Short run impacts of accountability on school quality. *American Economic Journal: Economic Policy 2*(4), 119–147.

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

Rothstein, J. (2010, February). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics 125*(2), 175–215.

Rouse, C. E., J. Hannaway, D. Goldhaber, and D. Figlio (2013). Feeling the florida heat? how low-performing schools respond to voucher and accountability pressure. *American Economic Journal: Economic Policy 5*(2), 251–281.

Scafidi, B., D. Sjoquist, and T. Stinebrickner (2007). Race, poverty, and teacher mobility. *Economics of Education Review 26*(2), 145–159.

Shipps, D. and M. White (2009). A new politics of the principalship? Accountability-driven change in New York City. *Peabody Journal of Education 84*(3), 350–373.

Shirrell, M. (2016). The effects of subgroup-specific accountability on teacher turnover and attrition. *Education Finance and Policy*.

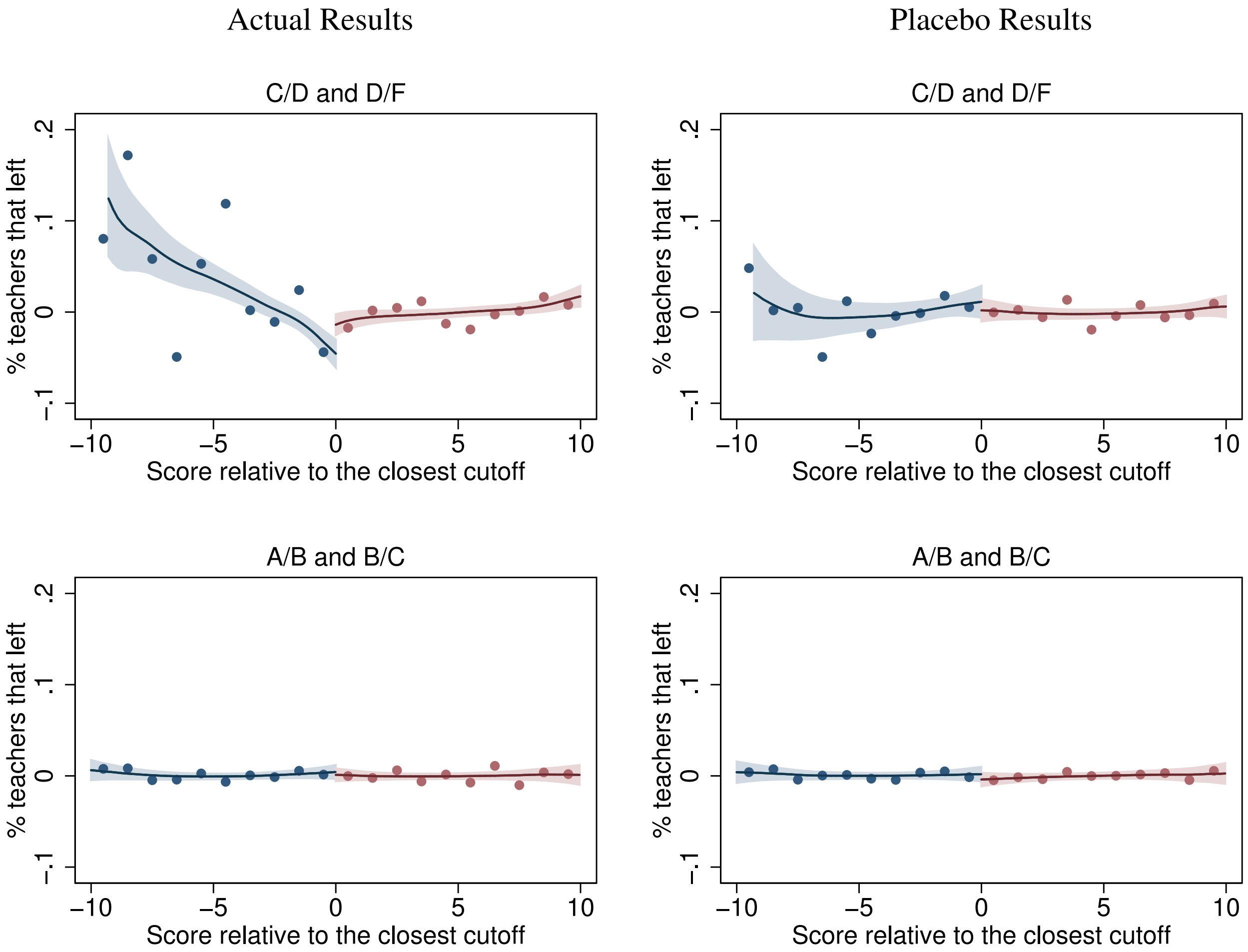
Smith, T. M. and R. M. Ingersoll (2004). What are the effects of induction and mentoring on beginning teacher turnover? *American educational research journal 41*(3), 681–714.

Springer, M. G. (2011). *New York City's school-wide bonus pay program: Early evidence from a randomized trial*. DIANE Publishing.

Venhorst, V., J. Van Dijk, and L. Van Wissen (2011). An analysis of trends in spatial mobility of dutch graduates. *Spatial Economic Analysis 6*(1), 57–82.

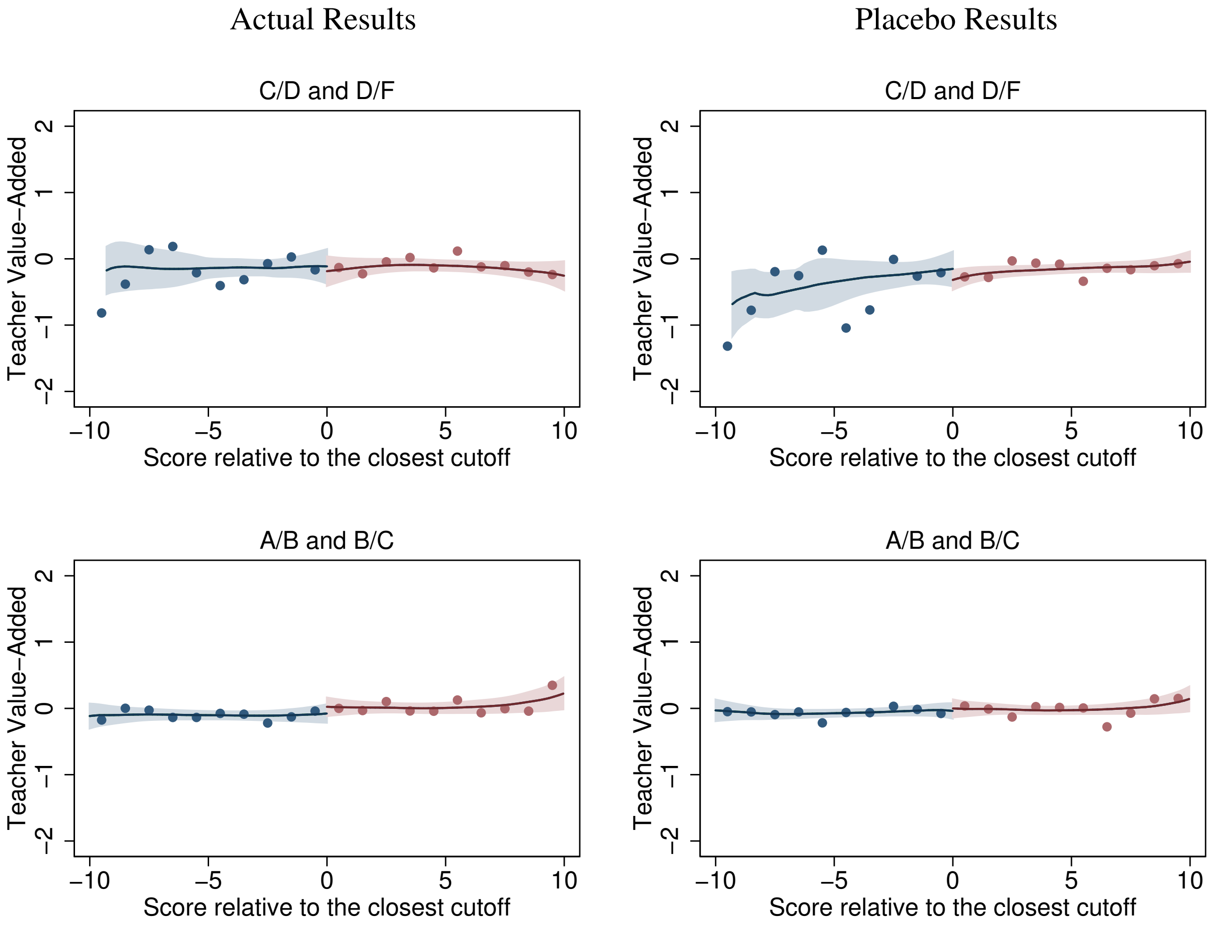
West, M. R. and P. E. Peterson (2006, March). The efficacy of choice threats within school accountability systems: Results from legislatively induced experiments. *The Economic Journal 116*, C46–C62.

Figure 1: Residual Turnover, by Accountability Score



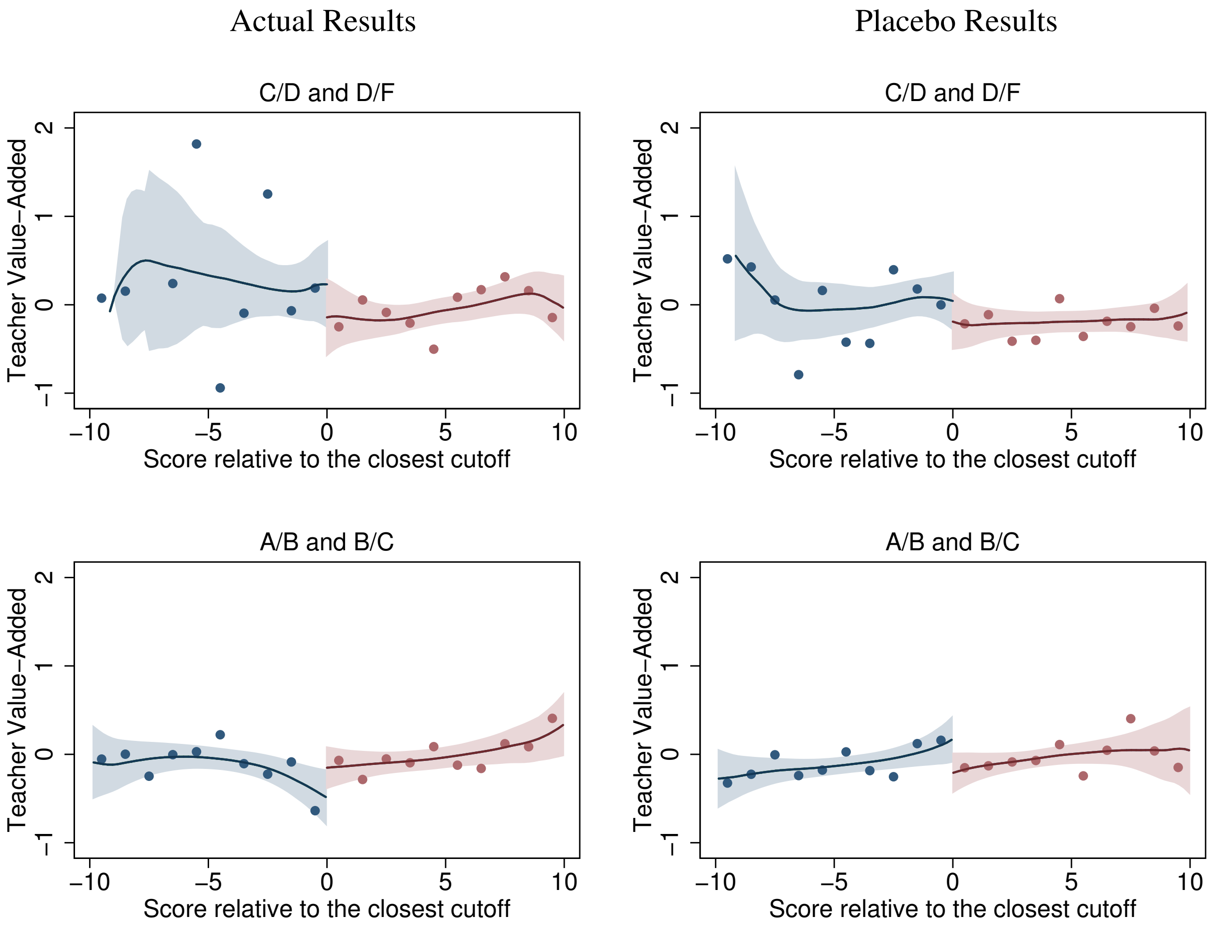
Notes. The left column plots the actual turnover results. The x-axes show schools' average accountability scores relative to the closest grade threshold (so the grade threshold is always displayed at 0). Each dot represents the average for all schools within a 1-point bandwidth on the x-axis. The y-axes show average residual turnover in the summer after the schools received their accountability grades. The right column shows placebo turnover results: there, the y-axes show residual turnover in the year *before* schools received grades. Residual turnover is calculated by regressing an indicator for leaving a school on a vector of covariates (see Table 2 notes for list of covariates). The lines correspond to local polynomial smooth plots, and the shaded areas represent their 95% confidence intervals.

Figure 2: Average Math Value-Added of Leavers, by Accountability Score



Notes. The left column plots the actual leaver quality results. The x-axes show the schools' average accountability scores relative to the grade threshold (so the grade threshold is always displayed at 0). The y-axes show the average value-added of leavers (i.e., of the teachers who left their schools in the summer after their schools received the accountability score and grade). The right column shows the placebo results: there, the y-axes show the average value-added of the teachers who left their schools the year *before* their schools received the accountability score and grade. The lines correspond to local polynomial smooth plots, and the shaded areas represent their 95% confidence intervals.

Figure 3: Average Math Value-Added of Joiners, by Accountability Score



Notes. The left column plots the actual joiner quality results. The x-axes show schools' average accountability scores relative to the closest grade threshold (so the grade threshold is always displayed at 0). Each dot represents the average for all schools within a 1-point bandwidth on the x axis. The y-axes show the average value-added of joiners (i.e., of the teachers who joined schools in the summer after their schools received the accountability score and grade). The right column shows the placebo results: there, the y-axes show the average value-added of the teachers who joined their schools the year *before* their schools received the accountability score and grade. The lines correspond to local polynomial smooth plots, and the shaded areas represent their 95% confidence intervals.

Table 1. Descriptive Statistics by Accountability Grade

		Accountability Grade				
	All Schools	A	B	C	D	F
Panel A: Teacher Characteristics						
% Teachers with Master's Degree	0.43	0.46	0.44	0.43	0.39	0.43
Teacher Experience (years)	9.79	9.90	9.94	9.71	9.26	9.30
Estimated Math Value-Added	-0.02	0.10	0.00	-0.07	-0.13	-0.28
Estimated ELA Value-Added	-0.03	0.07	-0.02	-0.07	-0.05	-0.21
Predicted Math Value-Added	0.12	0.16	0.13	0.11	0.06	0.09
Predicted ELA Value-Added	0.09	0.10	0.09	0.09	0.10	0.08
% Teachers that are:						
Female	0.83	0.84	0.83	0.82	0.80	0.81
Black	0.20	0.15	0.19	0.22	0.29	0.23
Non-Hispanic White	0.61	0.66	0.63	0.59	0.51	0.59
Hispanic	0.14	0.13	0.14	0.14	0.16	0.14
Asian	0.05	0.05	0.05	0.04	0.04	0.04
Turnover	0.11	0.10	0.11	0.12	0.15	0.19
Retirement	0.01	0.01	0.01	0.01	0.01	0.01
Intra-district transfers	0.04	0.03	0.03	0.04	0.06	0.10
Exited NYCDOE teacher files	0.07	0.06	0.07	0.07	0.08	0.08
Sample Size: Teacher-Year Observations (Base Sample)						
All	71,677	12,125	30,161	21,362	6,814	1,215
With Math Value-Added Data	19,092	3,154	8,119	5,639	1,885	295
Sample Size: Unique Teachers (Base Sample)						
All	50,616					
With Math Value-Added Data	13,202					
Panel B: School Characteristics						
Enrollment	809	662	702	697	600	521
% Students that are:						
Black	0.32	0.27	0.32	0.38	0.45	0.39
Non-Hispanic White	0.15	0.16	0.15	0.15	0.08	0.15
Hispanic	0.40	0.38	0.40	0.38	0.42	0.40
Asian	0.12	0.18	0.12	0.09	0.04	0.05
Immigrants	0.02	0.02	0.02	0.02	0.01	0.01
Components of Accountability Grades						
Environment Score	7.43	9.22	7.99	6.90	5.90	6.37
Performance Score	15.11	18.49	15.64	13.94	11.42	9.50
Progress Score	26.67	34.47	29.58	22.10	16.24	11.71
Additional Credit	2.18	3.32	2.40	1.22	0.75	0.26
Overall Score	51.40	65.51	55.62	44.16	34.33	27.85
Other characteristics						
NY School Bonus Program	0.15	0.14	0.17	0.15	0.18	0.23
Being phased out[a]	0.01	0.00	0.00	0.00	0.07	0.12
Not restructuring at baseline[b]	0.87	0.94	0.90	0.90	0.88	0.85
Sample Size: Schools						
Number of school-year observations	1,243	220	507	360	130	26
Number of unique schools	847					

Notes. Data come from the 2007-08 and 2008-09 school years in the New York City Department of Education. The accountability grade is the school report card grade that was received by the school during fall of the school year. Sample here limited to the sample used for the base turnover analysis: schools within a 5-point bandwidth of one of the grade thresholds.

a. Phase outs proxied for by schools that received accountability grades in one year but not the subsequent year.

b. Schools that were not undergoing restructuring prior to the institution of the accountability system (i.e., that were not in year 2+ of restructuring in the 2007-08 school year).

Table 2. Regression Discontinuity Estimates of the Effect of School Accountability Grades on Turnover

	Dependent Var.: Teacher Left School (Dummy)			
	Actual results (current year turnover)		Placebo results (previous year turnover)	
Sample:	Full sample	Non-restructuring sample	Full sample	Full sample, Year 1 only
	(1)	(2)	(3)	(4)
Bottom of the grade distribution (C/D and D/F)				
School received lower grade (dummy)	-0.029	-0.030	0.008	-0.003
	[0.013]**	[0.014]**	[0.016]	[0.020]
N	16,897	14,617	14,535	9,848
Dep. Var. Mean	0.13	0.12	0.14	0.14
Top of the grade distribution (A/B and B/C)				
School received lower grade (dummy)	0.005	0.003	0.002	0.004
	[0.007]	[0.007]	[0.007]	[0.011]
N	54,672	47,967	47,676	24,939
Dep. Var. Mean	0.11	0.11	0.12	0.13

Notes. Table presents regression discontinuity estimates of the effect of school accountability grades on teacher turnover. For the actual regressions, the dependent variable is teacher turnover, i.e., a dummy for the teacher leaving the school at the end of the school year. In the placebo regressions, the dependent variable is turnover in the year *prior* to the year the accountability grade was received. The sample is all teachers teaching in sample schools and each observation represents one teacher in a given year. Regressions use a bandwidth of 5 grade points. Standard errors are reported in brackets and clustered at the school level. All regressions include school covariates (i.e., controls for the average previous year's achievement; previous year's accountability score (second year only); the percent of students that are black, Hispanic, Asian, that receive free and reduced price lunch, and that are immigrants; fixed effects for school size; the percent of teachers that are black, Hispanic, or Asian; whether the school was in the NYSBP; and five-year average school turnover prior to the institution of accountability), teacher covariates (i.e., fixed effects for teacher experience and age, teacher education level, teacher gender, teacher race), and controls for the "RD running variable" (accountability score relative to the threshold) separately by (year)X(schooltype)X(received lower grade), as well as controls for (year)X(schooltype)X(received lower grade) and all of the main effects and lower-order interactions (e.g., year, schooltype, yearXschooltype, etc.). Data come from the 2008-09 and 2009-10 school years for the actual regressions and the 2007-08 school year for the placebo regressions. All data from the New York City Department of Education. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Table 3. RD Estimates of the Effect of Accountability Grades on the Characteristics of Leavers and Joiners

Panel A. Value-added

	Leavers				*Joiners*			
	Estimated math VA		Predicted math VA		Estimated math VA		Predicted math VA	
Sample:	Full	Non-restructuring	Full	Non-restructuring	Full	Non-restructuring	Full	Non-restructuring
Independent Variable = *Received **lower** grade at:*	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Bottom of the grade distribution								
D/F and C/D thresholds	-0.028	-0.150	0.124	0.021	0.869	0.887	0.164	0.214
	[0.21]	[0.24]	[0.10]	[0.12]	[0.34]**	[0.38]**	[0.11]	[0.11]**
N	479	385	2228	1765	126	113	1493	1339
Dep. Var. Mean	-0.12	-0.09	-0.07	-0.07	-0.12	-0.12	-0.05	-0.04
Top of the grade distribution								
B/C and A/B Thresholds	-0.130	-0.145	0.119	0.117	-0.392	-0.374	0.023	0.009
	[0.10]	[0.10]	[0.05]**	[0.05]**	[0.20]*	[0.22]*	[0.07]	[0.07]
N	1194	999	5916	5056	379	339	4251	3765
Dep. Var. Mean	-0.05	-0.05	-0.03	-0.01	-0.09	-0.11	-0.05	-0.04

Notes. Table presents regression discontinuity estimates of the effect of school accountability grades on the characteristics of the teachers that leave (leavers) and teachers that are hired by a given school (joiners). Specifically, each observation is a teacher in a given year. In Panel A, the dependent variable is the estimated math value-added (VA) of the teachers in the sample (columns (1)-(2) and (5)-(6)) or the predicted math VA (columns (3)-(4) and (7)-(8)). In Panel B, the dependent variable is a dummy for having VA data, the teacher's experience or education. The sample is all leaving (Panel A: columns (1)-(4), Panel B: columns (1)-(3)) or joining (Panel A: columns (5)-(8), Panel B: columns (4)-(6)) teachers and each observation represents one teacher in a given year. Regressions use a bandwidth of 5 grade points. Standard errors are reported in brackets and clustered at the school level. All regressions include controls for (year)X(schooltype)X(received lower grade) and all of the main effects and lower-order interactions (e.g., year, schooltype, yearXschooltype, etc.), as well as school controls (which include controls for the average previous year's achievement; previous year's accountability score (second year only); the percent of students that are black, Hispanic, Asian, that receive free and reduced price lunch, and that are immigrants; fixed effects for school size; whether the school was in the NYSBP; and five-year average school turnover prior to the institution of accountability. Data come from the 2008-09 and 2009-10 school years in the New York City Department of Education for the actual regressions and the 2007-08 and 2008-09 school years for the placebo regressions, with the report card grades used being the report card that was received by the school during fall of the school year. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Panel B. Other teacher characteristics

Sample:	Leavers			Joiners		
Dependent Variable:	Has estimated VA data	At least 4 years experience	Has masters	Has estimated VA data	At least 4 years experience	Has masters
Independent Variable = Received ***lower*** *grade at:*	(1)	(2)	(3)	(4)	(5)	(6)
<u>Bottom of the grade distribution</u>						
D/F and C/D thresholds	0.060	-0.047	-0.051	0.003	-0.048	-0.031
	[0.04]	[0.05]	[0.05]	[0.03]	[0.06]	[0.04]
N	2228	2228	2228	1493	1493	1493
Dep. Var. Mean	0.21	0.51	0.33	0.08	0.28	0.23
<u>Top of the grade distribution</u>						
B/C and A/B Thresholds	-0.009	0.045	-0.012	-0.025	-0.001	-0.013
	[0.02]	[0.03]	[0.03]	[0.02]	[0.03]	[0.03]
N	5916	5916	5916	4251	4251	4251
Dep. Var. Mean	0.20	0.54	0.36	0.09	0.26	0.21

<u>Notes</u>. Table presents regression discontinuity estimates of the effect of school accountability grades on the characteristics of the teachers that leave (leavers) and teachers that are hired by a given school (joiners). Specifically, each observation is a teacher in a given year. In Panel A, the dependent variable is the estimated math value-added (VA) of the teachers in the sample (columns (1)-(2) and (5)-(6)) or the predicted math VA (columns (3)-(4) and (7)-(8)). In Panel B, the dependent variable is a dummy for having VA data, the teacher's experience or education. The sample is all leaving (Panel A: columns (1)-(4), Panel B: columns (1)-(3)) or joining (Panel A: columns (5)-(8), Panel B: columns (4)-(6)) teachers and each observation represents one teacher in a given year. Regressions use a bandwidth of 5 grade points. Standard errors are reported in brackets and clustered at the school level. All regressions include controls for (year)X(schooltype)X(received lower grade) and all of the main effects and lower-order interactions (e.g., year, schooltype, yearXschooltype, etc.), as well as school controls (which include controls for the average previous year's achievement; previous year's accountability score (second year only); the percent of students that are black, Hispanic, Asian, that receive free and reduced price lunch, and that are immigrants; fixed effects for school size; whether the school was in the NYSBP; and five-year average school turnover prior to the institution of accountability. Data come from the 2008-09 and 2009-10 school years in the New York City Department of Education for the actual regressions and the 2007-08 and 2008-09 school years for the placebo regressions, with the report card grades used being the report card that was received by the school during fall of the school year. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Table 4. Heterogeneity in the turnover effects by teacher characteristics

	Teacher characteristic used for heterogeneity:			
Sample:	Above-median estimated VA	Above-median predicted VA	At least 4 years experience	Has masters
	(1)	(2)	(3)	(4)
Bottom of the grade distribution (C/D and D/F)				
School received lower grade (dummy)	0.019	-0.025	-0.027	-0.030
	[0.023]	[0.020]	[0.030]	[0.019]
School rec'd lower grade X (Teacher characteristic)	-0.061	-0.008	0.000	0.006
	[.034]*	[.022]	[.03]	[.022]
N	4,524	16,897	16,897	16,897
Top of the grade distribution (A/B and B/C)				
School received lower grade (dummy)	0.020	0.004	-0.003	0.012
	[0.014]	[0.010]	[0.015]	[0.009]
School rec'd lower grade X (Teacher characteristic)	-0.026	0.004	0.011	-0.016
	[.018]	[.012]	[.016]	[.012]
N	14,544	54,672	54,672	54,672

Notes. Table presents regression discontinuity estimates of heterogeneity in the turnover effects by teacher characteristics. All regressions include the standard "RD controls" (i.e., a control for (year)X(schooltype)X(received lower grade) and a dummy for received lower grade interacted with the teacher characteristic listed at the top of the column, as a well as a control for the teacher characteristic itself. All results shown with the full sample of schools but results consistent in the non-restructuring sample. The dependent variable is teacher turnover, i.e., a dummy for the teacher leaving the school at the end of the school year. The sample is all teachers teaching in sample schools and each observation represents one teacher in a given year. Regressions use a bandwidth of 5 grade points. Standard errors are reported in brackets and clustered at the school level. All regressions include school covariates (i.e., controls for the average previous year's achievement; previous year's accountability score (second year only); the percent of students that are black, Hispanic, Asian, that receive free and reduced price lunch, and that are immigrants; fixed effects for school size; the percent of teachers that are black, Hispanic, or Asian; whether the school was in the NYSBP; and five-year average school turnover prior to the institution of accountability), teacher covariates (i.e., fixed effects for teacher experience and age, teacher education level, teacher gender, and teacher race), and a year dummy. Data come from the 2008-09 and 2009-10 school years for the actual regressions and the 2007-08 school year for the placebo regressions. All data from the New York City Department of Education. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Table 5. Regression Discontinuity Estimates of the Effects of School Grades on Teacher Survey Responses

Panel A. Teacher survey responses: Summary measures

	Dependent Variable = Standardized index of teacher survey responses about:						
Independent Var. =	Principal leadership	Setting high expectations for students	School course offerings	Teacher collaboration	Professional development	Parent interaction	Safety
*Received **lower** grade at the:*	(1)	(2)	(3)	(4)	(5)	(6)	(7)
<u>Bottom of the grade distribution</u>							
D/F or C/D thresholds	0.27	0.19	-0.01	0.08	-0.06	0.11	-0.09
	[0.16]*	[0.14]	[0.10]	[0.15]	[0.16]	[0.11]	[0.05]*
N	309	309	309	309	309	309	309
<u>Top of the grade distribution</u>							
B/C or A/B Thresholds	-0.16	-0.08	-0.16	0.04	0.01	-0.08	0.04
	[0.10]	[0.09]	[0.05]***	[0.08]	[0.09]	[0.07]	[0.03]
N	932	932	932	931	932	932	932

Panel B. Teacher survey responses: Detailed survey responses for principal leadership questions

	Dependent Variable = Level of teacher agreement with the following statements:							
Independent Var. =	School leaders communicate a clear vision for the school	School leaders let staff know what is expected of them	School leaders encourage open communicati on on important school issues	Curriculum, instruction, and assessment are aligned w/in and across grade levels at the school	The principal places the learning needs of children ahead of other interests	The principal is an effective manager who makes the school run smoothly	I trust the principal at his/her word	To what extent do you feel supported by the principal?
*Received **lower** grade at the:*	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
<u>Bottom of the grade distribution</u>								
D/F or C/D thresholds	0.26	0.29	0.25	0.30	0.24	0.32	0.28	0.35
	[0.19]	[0.20]	[0.19]	[0.15]*	[0.19]	[0.19]*	[0.18]	[0.18]*
N	309	309	309	309	309	309	309	309
<u>Top of the grade distribution</u>								
B/C or A/B Thresholds	-0.18	-0.15	-0.18	-0.08	-0.21	-0.24	-0.22	-0.21
	[0.11]*	[0.11]	[0.11]	[0.09]	[0.11]*	[0.11]**	[0.11]*	[0.11]*
N	932	932	932	932	932	932	932	932

<u>Notes</u>. Table presents regression discontinuity estimates of the effect of school accountability grades on teacher survey responses. The dependent variables are school-average responses from teacher survey questions, expressed in standard deviations of the response distribution, with all questions normalized so that more positive is better. For panel A, the variables are indexes that average the responses across several sub-questions in a given area; for panel B, the variables are responses to the individual sub-questions in the principal leadership index Each observation represents one teacher in a given year. Regressions use a bandwidth of 5 grade points. Standard errors are reported in brackets and clustered at the school level. All regressions include controls for (year)X(schooltype)X(received lower grade) and all of the main effects and lower-order interactions (e.g., year, schooltype, yearXschooltype, etc.) and school covariates, which include controls for the average previous year's achievement; previous year survey responses; previous year's accountability score (second year only); the percent of students that are black, Hispanic, Asian, that receive free and reduced price lunch, and that are immigrants; the percent of teachers that are black, Hispanic, or Asian; fixed effects for school size; whether the school was in the NYSBP; and five-year average school turnover prior to the institution of accountability. Regressions are weighted by the number of teachers who filled out the surveys. Data come from the 2008-09 and 2009-10 school years. All data from the New York City Department of Education. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Table 6. Correlations Between Labor Market Outcomes and School Performance or Leadership

	Turnover Results		Joiner VA Results	
Dependent Variable:	*Teacher Left School*		*Estimated Math Value-Added*	
Sample:	*Incumbents*		*Joiners*	
Independent Variable:	(1)	(2)	(3)	(4)
School achievement (standardized)	-0.0205*** [0.00739]		0.480*** [0.153]	
Principal leadership rating (standardized)		-0.0133*** [0.00176]		0.0445 [0.0401]
N	111,926	111,926	801	801

Notes. Table presents estimates of the correlation between labor market outcomes (specifically, turnover at the end of the year (columns (1)-(2)) or the quality of joiners hired in the subsequent year (column (3)-(4)), and either school achievement (columns (1) and (3)) or school principal leadership ratings (columns (2) and (4)). Achievement is the average between math and English; principal leadership ratings come from teacher surveys. Both achievement and principal leadership are expressed in standard deviations. Sample covers 2008-2009 and includes all schools that received accountability grades. For columns (1)-(2), the dependent variable is an indicator for whether a teacher stopped teaching at the school; the sample is all teachers teaching in sample schools, and each observation represents one teacher in a given year. For columns (3)-(4), the dependent variable is a teacher’s math value-added; the sample is all joiners to a school at the end of the year, and each observation represents one teacher in a given year. Standard errors are reported in brackets and clustered at the school level. Columns (1) and (3) include school covariates (i.e., controls for the average previous year's achievement; the previous year's accountability score (second year only); the percent of students that are black, Hispanic, Asian, that receive free and reduced price lunch, and that are immigrants; fixed effects for school size; whether the school was in the NYSBP; and five-year average school turnover prior to the institution of accountability), teacher covariates (i.e., fixed effects for teacher experience and age, teacher education level, teacher gender, and teacher race), and an accountability score control. Columns (2) and (4) include the same covariates except the set of teacher controls, and also add in previous year principal leadership controls. All data from the New York City Department of Education. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Table 7. Heterogeneity in the RD effects by principal quality

Dependent Variable:	Turnover Dummy	End-of-year principal leadership survey ratings	Joiner quality
	(1)	(2)	(3)
Bottom of the grade distribution (C/D and D/F)			
Lower grade	-0.012	0.080	0.998
	[0.019]	[0.215]	[0.70]
(Lower grade) X (Above-median baseline principal leadership)	-0.060	0.154	-0.331
	[.029]**	[.33]	[0.97]
N	16,833	308	124
Top of the grade distribution (A/B and B/C)			
Lower grade	-0.107	-0.106	-0.359
	[0.165]	[0.134]	[0.29]
(Lower grade) X (Above-median baseline principal leadership)	0.052	0.139	-0.205
	[.206]	[.191]	[.451]
N	927	927	378
Sample	Teachers	Schools	Joiners to school

Notes. Table presents regression discontinuity estimates of heterogeneity in the turnover effects by principal quality. Regressions use a bandwidth of 5 grade points. Standard errors are reported in brackets and clustered at the school level. All regressions include the standard "RD controls" (i.e., controls for (year)X(schooltype)X(received lower grade) and all of the main effects and lower-order interactions (e.g., year, schooltype, yearXschooltype, etc.)) as well as a dummy for received lower grade interacted with the dummy for above-median baseline principal leadership, and a control for above-median baseline principal leadership. Baseline principal leadership is the leadership rating for the principal of that school in the previous year. All regressions include school covariates (i.e., controls for the average previous year's achievement; previous year's accountability score (second year only); the percent of students that are black, Hispanic, Asian, that receive free and reduced price lunch, and that are immigrants; the percent of teachers that are black, Hispanic, or Asian; fixed effects for school size; whether the school was in the NYSBP; and five-year average school turnover prior to the institution of accountability) and a year dummy. Column (1) also includes teacher covariates (i.e., fixed effects for teacher experience and age, teacher education level, teacher gender, and teacher race). Data come from the 2008-09 and 2009-10 school years for the actual regressions and the 2007-08 school year for the placebo regressions. All data from the New York City Department of Education. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Table 8. Heterogeneity in the Regression Discontinuity Turnover Estimates by Teacher Destination

	Teacher turnover and transfer application submission							*School transfer applications received*
	Payroll data				Open market transfer data			Open market transfer data
Dependent Variable:	Left	Left NYCDOE Classrooms	Retire	Transfer	Submitted application to transfer	Percent applications leading to transfers in open market data	Percent applications leading to transfers in payroll data	Transfer applications received (as % current staff size)
Independent Var. = *School received* ***lower*** *grade at the:*	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Bottom of the grade distribution								
D/F or C/D thresholds	-0.029	-0.023	-0.003	-0.003	0.006	0.062	0.048	0.08
	[0.013]**	[0.008]***	[0.004]	[0.010]	[0.014]	[0.062]	[0.067]	[0.206]
N	16,897	16,897	16,897	16,897	16,897	1,604	1,604	309
Dep. Var. Mean	0.13	0.07	0.01	0.05	0.09	0.35	0.44	0.92
Top of the grade distribution								
B/C or A/B Thresholds (grouped)	0.005	0.003	0.001	0.000	0.011	-0.014	-0.003	0.08
	[0.007]	[0.005]	[0.002]	[0.005]	[0.006]*	[0.032]	[0.035]	[0.108]
N	54,672	54,672	54,672	54,672	54,672	3,457	3,457	932
Dep. Var. Mean	0.11	0.07	0.01	0.03	0.06	0.34	0.42	0.98
Each observation is a	Teacher	Teacher	Teacher	Teacher	Teacher	Teacher	Teacher	School
What proportionate coefficient would be for the grouped C/D and D/F thresholds	-0.029	-0.015	-0.002	-0.011				

Notes. Table presents regression discontinuity estimates of the effect of school accountability grades on teacher turnover by teacher destination. Columns (1) through (4) use the payroll data to look at turnovers by destination: Column (1) has overall turnover, and columns (2)-(4) break up departures between the three main ways teachers can leave the school (retirements, transfers, and stopping working in NYCDOE classrooms). Columns (5)-(8) use the open market transfer data on submitted transfer applications to look at the effects on whether teachers applied to transfer; whether their transfer application was listed as leading to a successul transfer in the open market transfer data (col. (6)); whether that application was associated with a transfer in the payroll data (col. (7)); and on the number of transfer applications received by lower-graded schools (col. (8)). The sample is all teachers teaching in sample schools during the 2008-2009 and 2009-2010 school years and each observation represents one teacher in a given year, except for column (8), where each observation is a school and the regression is weighted by the number of teachers. Regressions use a bandwidth of 5 grade points. Standard errors are reported in brackets and clustered at the school level. All regressions include school controls (which include controls for the average previous year's achievement; the previous year's accountability score (second year only); the percent of students that are black, Hispanic, Asian, that receive free and reduced price lunch, and that are immigrants; fixed effects for school size; whether the school was in the NYSBP; and five-year average school turnover prior to the institution of accountability), teacher covariates (which include fixed effects for teacher experience and age, teacher education level, teacher gender, and teacher race) and controls the "RD running variable" (accountability score relative to the threshold) separately by (year)X(schooltype)X(received lower grade), as well as controls for (year)X(schooltype)X(received lower grade) and all of the main effects and lower-order interactions (e.g., year, schooltype, yearXschooltype, etc.), except column (8) which excludes the set of teacher controls. Data come from the 2008-09 and 2009-10 school years in the New York City Department of Education.

* Significant at 10%; ** significant at 5%; *** significant at 1%.

Table 9. Robustness of the Regression Discontinuity Turnover Estimates

Specification:	Local linear, different bandwidths and controls						Quadratic	Cubic	Detailed score control	Local linear, school level, weighted
Bandwidth:	5 (base)	5 (base)	2.5 (base/2)	4 (base-1)	6 (base+1)	10 (baseX2)	10	10	5	5
Ind. Var. = Received ***lower*** *grade at the:*	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Bottom of the grade distribution										
D/F or C/D (grouped)	-0.018	-0.029	-0.033	-0.009	-0.028	-0.022	-0.030	-0.060	-0.017	-0.030
	[0.017]	[0.013]**	[0.017]*	[0.014]	[0.014]**	[0.014]	[0.020]	[0.023]***	[0.017]	[0.015]**
N	16,897	16,897	9,882	13,706	18,521	28,452	28,452	28,452	16,897	309
P-Value: Spec Test						0.24	0.53	0.21	0.44	0.34
Top of the grade distribution										
B/C or A/B (grouped)	0.008	0.005	0.001	0.003	0.002	0.000	0.003	-0.002	0.001	0.005
	[0.008]	[0.007]	[0.010]	[0.007]	[0.006]	[0.005]	[0.007]	[0.010]	[0.003]	[0.007]
N	54,672	54,672	27,627	43,511	64,091	88,132	88,132	88,132	54,672	932
P-Value: Spec Test						0.80	0.78	0.67	0.54	0.86
School covariates?	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Teacher covariates?	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No

Notes. Table presents regression discontinuity estimates of the effect of school accountability grades on teacher turnover. The dependent variable is an indicator for whether a teacher stopped teaching at the school in the summer after the accountability grade was received. The sample is all teachers teaching in sample schools during the 2008-2009 and 2009-2010 school years, and each observation represents one teacher in a given year (columns (1)-(9)) or one school in a given year (column (10)). Column (10) displays the results from column (2) collapsed to the school level instead of the teacher level, with the estimates weighted by the number of teachers at a school. Regressions control for the "RD running variable" (accountability score relative to the threshold) separately by (year)X(schooltype)X(received lower grade), as well as controls for (year)X(schooltype)X(received lower grade) and all of the main effects and lower-order interactions (e.g., year, schooltype, yearXschooltype, etc.). The control function is linear for columns (1)-(6) and (10), quadratic for column (7), and cubic for column (8). School controls include controls for the average previous year's achievement; the previous year's accountability score (second year only); the percent of students that are black, Hispanic, Asian, that receive free and reduced price lunch, and that are immigrants; whether the school was in the NYSBP; fixed effects for school size; and five-year average school turnover prior to the institution of accountability. Teacher covariates include fixed effects for teacher experience and age, teacher education level, teacher gender, and teacher race. Data come from the 2008-09 and 2009-10 school years in the New York City Department of Education. The specification test tests for discontinuities in the regression function other than at the specified threshold (specifically, at all points that are a multiple of one point from the true threshold); the p-value represents the p-value from a joint test that there are no discontinuities at any other points. * Significant at 10%; ** significant at 5%; *** significant at 1%.

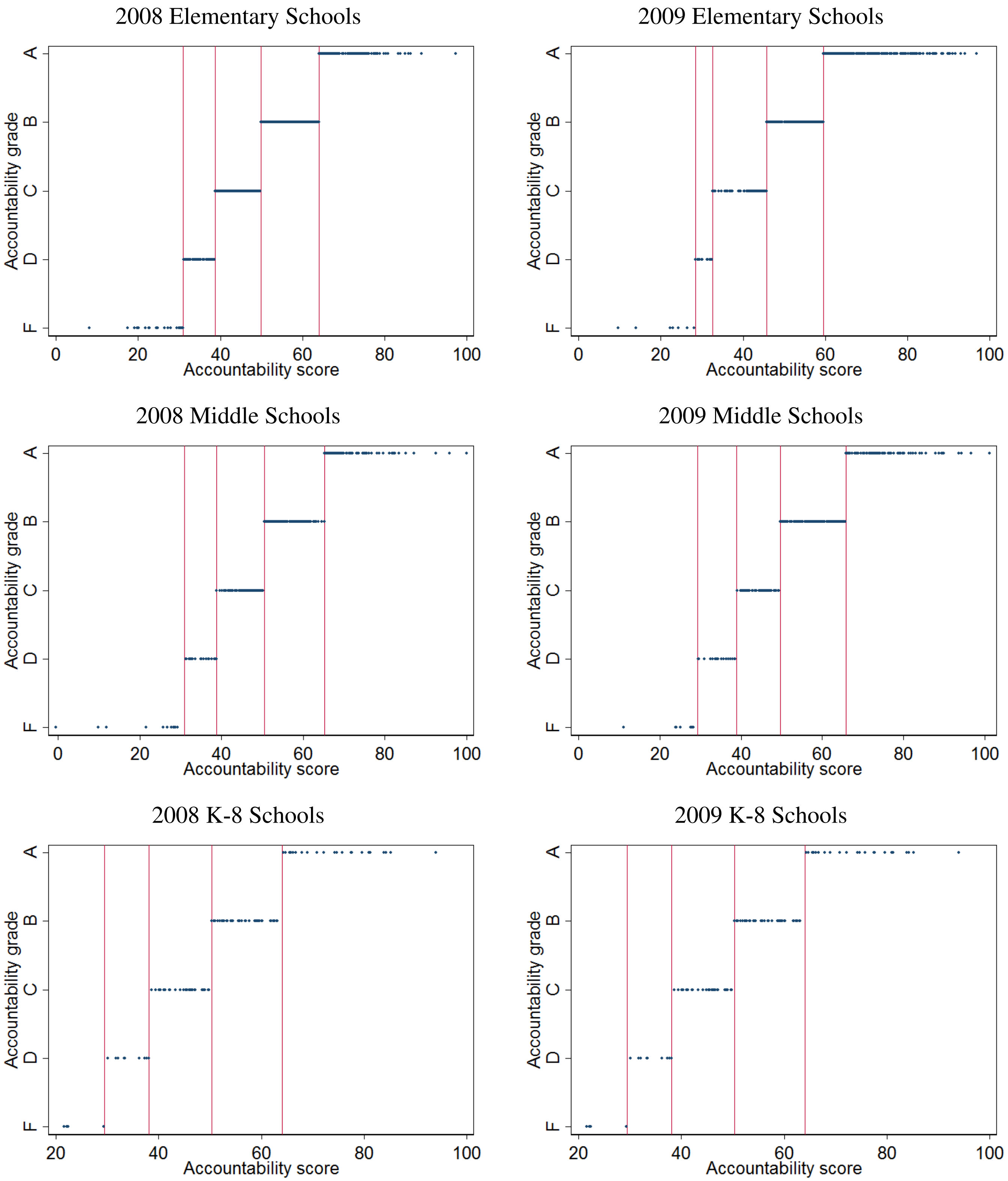
Table 10. Robustness of the Regression Discontinuity Joiner Quality Estimates

Dependent Variable:	Estimated Math VA									
Specification:	Local linear, different bandwidths and controls						Quadratic	Cubic	Detailed Score Control	Local linear, school level
Bandwidth:	5 (base)	5 (base)	2.5 (base/2)	4 (base-1)	6 (base+1)	10 (baseX2)	10	10	5	5
*Ind. Var. = Received **lower** grade at the:*	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Bottom of the grade distribution										
D/F or C/D (grouped)	0.789	0.869	0.411	0.757	0.888	0.465	0.494	-0.049	0.474	0.869
	[0.34]**	[0.34]**	[0.53]	[0.38]**	[0.31]***	[0.27]*	[0.38]	[0.53]	[0.19]**	[0.36]**
N	126	126	66	100	146	216	216	216	126	94
P-Value: Spec Test						0.24	0.01	0.58	0.57	0.30
Top of the grade distribution										
B/C or A/B (grouped)	-0.365	-0.392	-0.434	-0.521	-0.316	-0.077	-0.398	-0.648	-0.070	-0.392
	[0.20]*	[0.20]*	[0.31]	[0.22]**	[0.18]*	[0.14]	[0.23]*	[0.30]**	[0.10]	[0.21]*
N	379	379	166	288	452	629	629	629	379	269
P-Value: Spec Test						0.72	0.26	0.13	0.19	0.73
School covariates?	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes. Table presents regression discontinuity estimates of the effect of school accountability grades on the quality of joiners hired in the subsequent year. The dependent variable is a teacher's math value-added; the sample is all joiners to a school at the end of the year, and each observation represents one teacher in a given year, except for column (10) which collapses the data to the school level and estimates the regression weighted by the number of teachers. Regressions control for the "RD running variable" (accountability score relative to the threshold) separately by (year)X(schooltype)X(received lower grade), as well as controls for (year)X(schooltype)X(received lower grade) and all of the main effects and lower-order interactions (e.g., year, schooltype, yearXschooltype, etc.). The control function is linear for columns (1)-(6) and (10), quadratic for column (7), cubic for column (8), and controls for the factors underlying the accountability score in column (9). School covariates include controls for the average previous year's achievement; the previous year's accountability score (second year only); the percent of students that are black, Hispanic, Asian, that receive free and reduced price lunch, and that are immigrants; whether the school was in the NYSBP; fixed effects for school size; and five-year average school turnover prior to the institution of accountability. Data come from the 2008-09 and 2009-10 school years in the New York City Department of Education. The specification test tests for discontinuities in the regression function other than at the specified threshold (specifically, at all points that are a multiple of one point from the true threshold); the p-value represents the p-value from a joint test that there are no discontinuities at any other points. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Appendix Figure A1: Relationship between Accountability Scores and Accountability Grades

(Each year and schooltype shown separately)


2008 Elementary Schools

2009 Elementary Schools

2008 Middle Schools

2009 Middle Schools

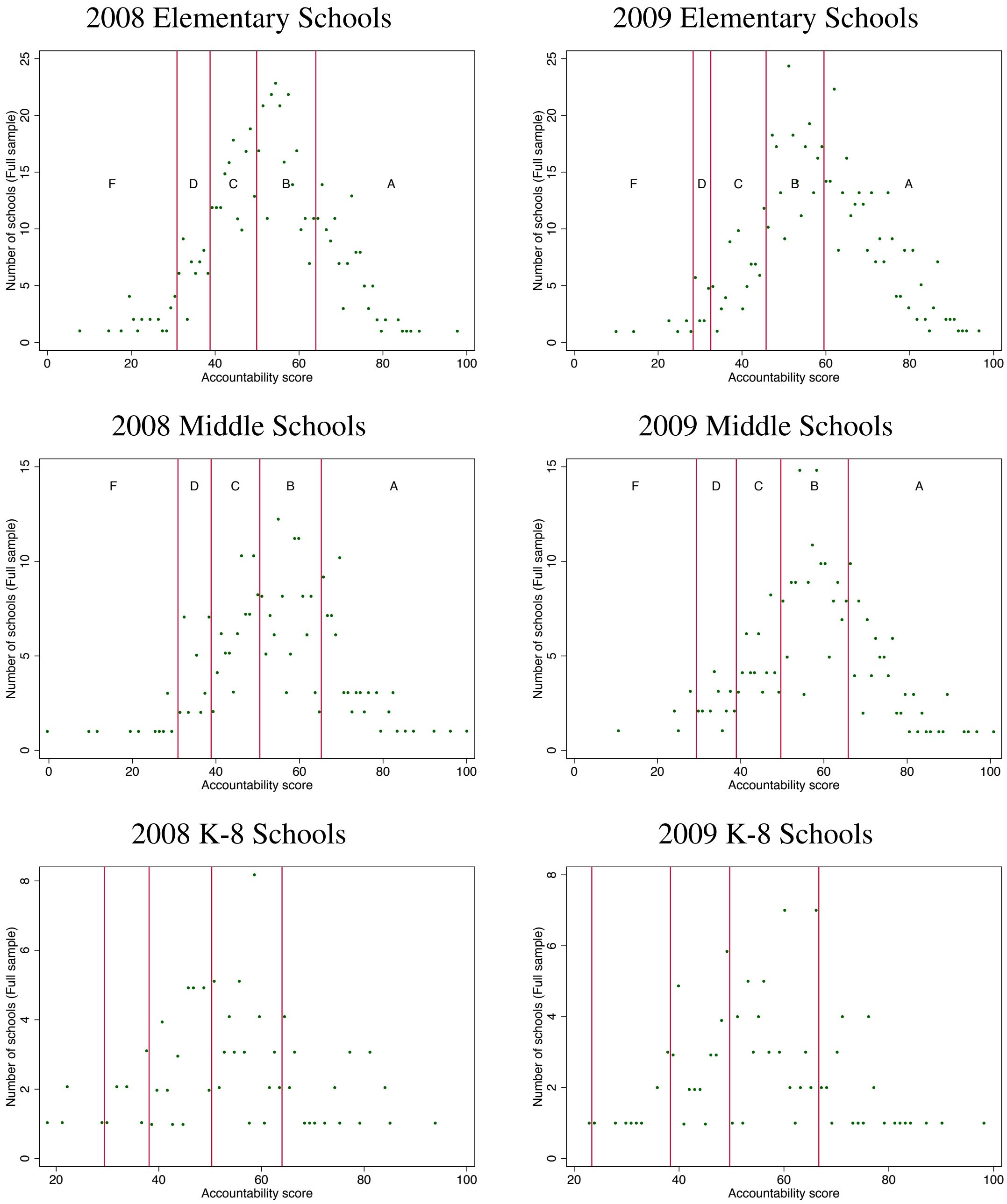
2008 K-8 Schools

2009 K-8 Schools

Notes. For each year and school type, the figures plot the accountability grade received by a school as a function of the underlying accountability score.

Appendix Figure A2: Density of Schools Near Grade Thresholds

(Each year and schooltype shown separately)



2008 Elementary Schools

2009 Elementary Schools

2008 Middle Schools

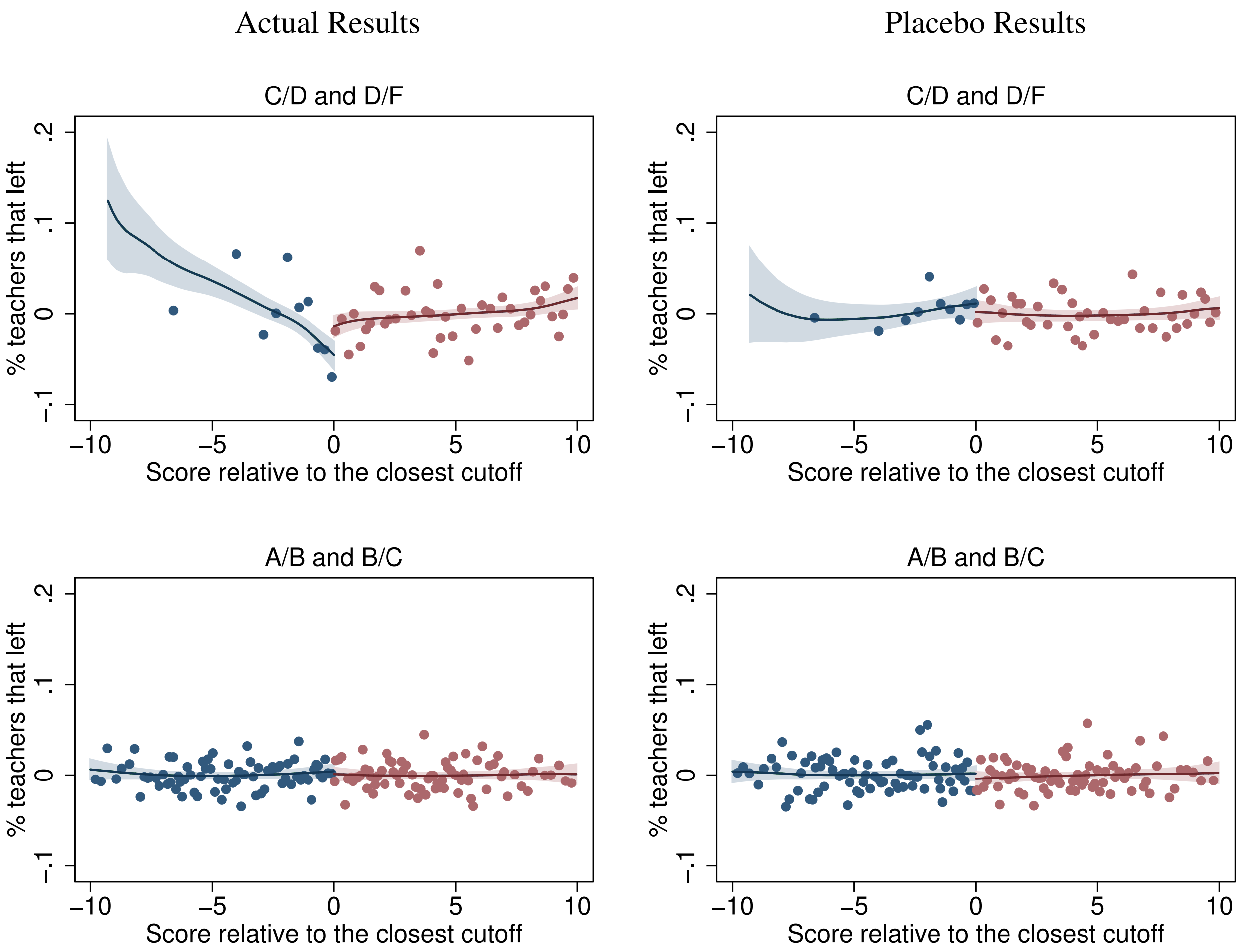
2009 Middle Schools

2008 K-8 Schools

2009 K-8 Schools

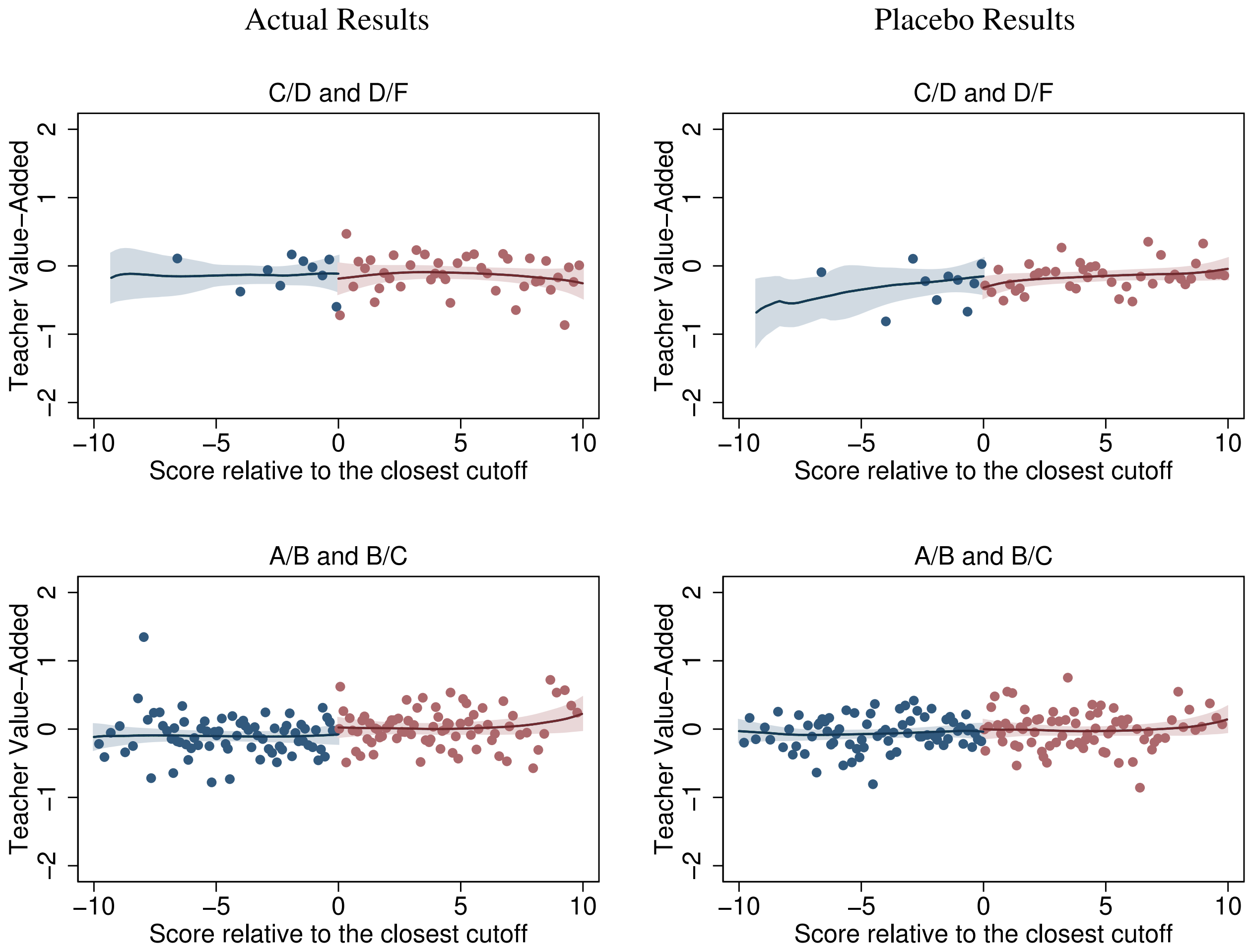
Notes. For each year and school type, the figures plot the number of schools with a given accountability score (specifically, the y-axis shows the number of schools within a 0.5 point bandwidth of the accountability score displayed on the X-axis). The red lines show the 4 grade thresholds (A/B, B/C, C/D, and D/F). Evidence of heaping directly adjacent to the grade thresholds line would be a violation of the regression discontinuity identification assumptions.

Appendix Figure A3: Residual Turnover, by Accountability Score
(version with a fixed number of schools per dot)



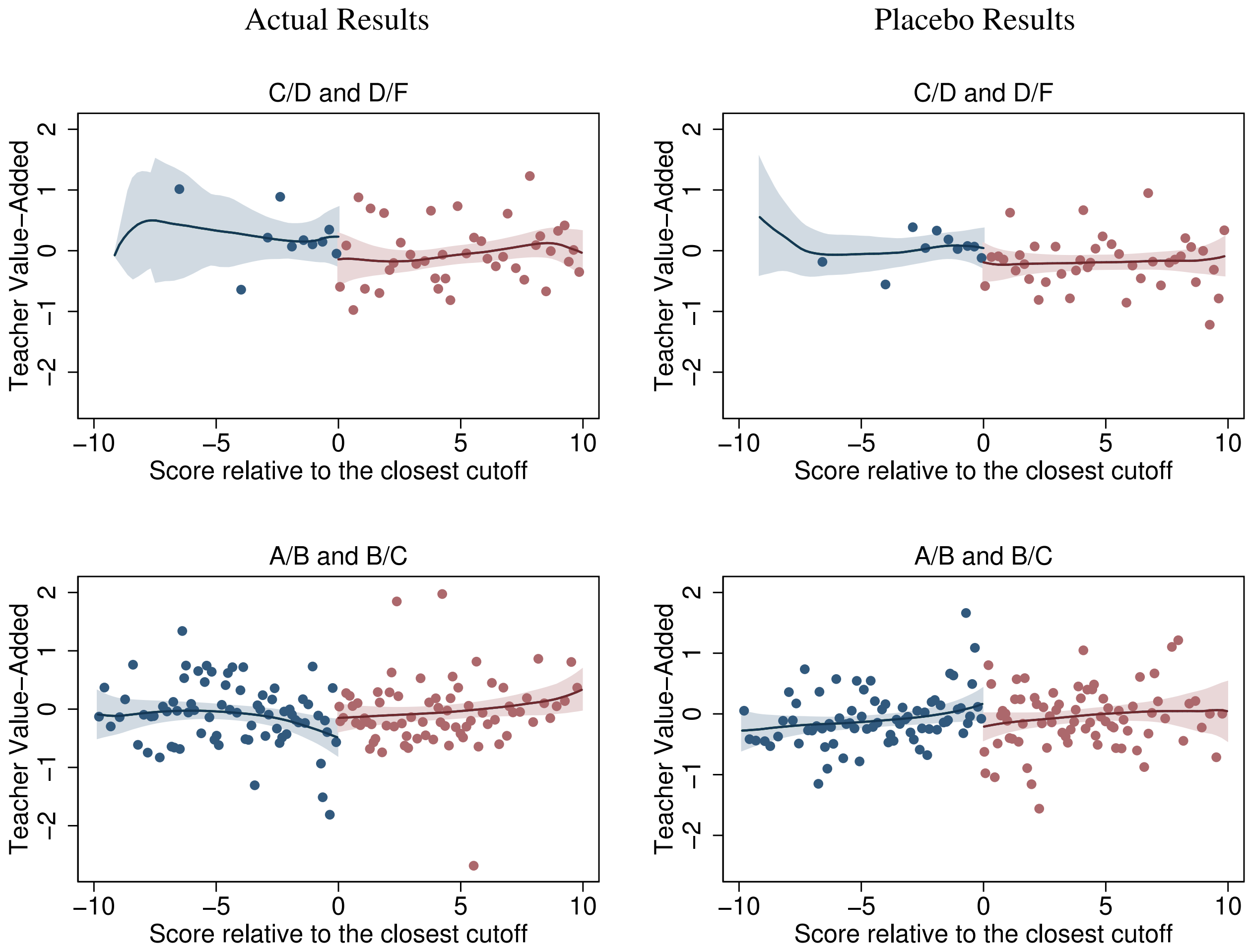
Notes. Each dot represents 10 schools. The left column plots the actual turnover results. The x-axes show the schools' average accountability scores relative to the closest grade threshold (so the grade threshold is always displayed at 0) and the y-axes show average residual turnover in the summer after the schools received their accountability grades. The right column shows placebo turnover results: there, the y-axes show residual turnover in the year *before* schools received grades (placebo uses data from the first year only). Residual turnover is calculated by regressing an indicator for leaving a school on a vector of covariates (see Table 2 notes for list of covariates). The lines correspond to local polynomial smooth plots, and the shaded areas represent their 95% confidence intervals.

Appendix Figure A4: Average Math Value-Added of Leavers, by Accountability Score (version with a fixed number of schools per dot)



Notes. Each dot represents 10 schools. The left column plots the actual leaver quality results. The x-axes show the schools' average accountability scores relative to the closest grade threshold (so the grade threshold is always displayed at 0). The y-axes show the average value-added of leavers (i.e., of the teachers who left their schools in the summer after their schools received the accountability score and grade). The right panel has the placebo results: there, the y-axes show the average value-added of the teachers who left their schools the year *before* their schools received the accountability score and grade (placebo uses data from the first year only). The lines correspond to local polynomial smooth plots, and the shaded areas represent their 95% confidence intervals.

Appendix Figure A5: Average Math Value-Added of Joiners, by Accountability Score (version with a fixed number of schools per dot)



Notes. Each dot represents 10 schools. The left panel plots the actual joiner quality results. The x-axes show the schools' average accountability scores relative to the closest grade threshold (so the grade threshold is always displayed at 0). The y-axes show the average value-added of joiners (i.e., of the teachers who joined schools in the summer after their schools received the accountability score and grade). The right panel has the placebo results: there, the y-axes show the average value-added of the teachers who joined their schools the year *before* their schools received the accountability score and grade (placebo uses data from the first year only). The lines correspond to local polynomial smooth plots, and the shaded areas represent their 95% confidence intervals.

Appendix Table A1. McCrary test

	Number of schools in a bin	
Dependent Variable:		
Binwidth:	0.5	0.1
Independent Variable:	(1)	(2)
Bottom of the grade distribution (C/D and D/F)		
School received lower grade (dummy)	-2.585	-0.561
	[3.326]	[0.845]
N	20	100
Top of the grade distribution (A/B and B/C)		
School received lower grade (dummy)	-3.436	-0.807
	[4.865]	[1.281]
N	20	100

Notes. Table shows the results of the McCrary test for gaming. The dependent variable is the number of schools in a bin, using a binwidth of 0.5 (column (1)) or 0.1 point (column (2)). Each column contains two cells containing the results of two separate regressions (one with schools at the bottom of the grade distribution, one with those at the top). Regressions use a bandwidth of 5 grade points. All regressions include a control for the score relative to the closest threshold, allowed to vary on either side of the threshold. Robust standard errors are reported in brackets. Data come from the 2007-08 and 2008-09 school years in the New York City Department of Education, using the report card grade that was received by the school during fall of the school year. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Appendix Table A2. No Selection Into the Analysis Sample

Dependent Variable:	School closed	School is missing teacher data	School being phased out	School not restructuring at baseline	School was part of New York School Bonus Program (NYSBP)
Independent Variable = *School received **lower** grade at the:*	(1)	(2)	(3)	(4)	(5)
Bottom of the grade distribution					
D/F or C/D thresholds (grouped)	0.000	-0.023	0.011	-0.107	0.000
	[0.000]	[0.020]	[0.032]	[0.080]	[0.000]
N	312	312	312	312	312
Dep. Var. Mean	0.003	0.006	0.038	0.901	0.173
Top of the grade distribution					
B/C or A/B Thresholds (grouped)	n/a	n/a	0.007	-0.014	0.000
			[0.005]	[0.031]	[0.000]
N	932	932	932	932	932
Dep. Var. Mean	0.000	0.000	0.002	0.905	0.157

Notes. The table presents regression discontinuity estimates of the effect of school accountability grades on school characteristics. Regressions use a bandwidth of 5 grade points. Standard errors are reported in brackets and clustered at the school level. Each observation is a school in a given year. Cols (1) and (2): The sample is all schools receiving accountability grades, and the dependent variable is an indicator that the school is excluded from the analysis sample because it closed or was missing teacher data (4 schools total). Cols (3)-(6): The sample is all non-excluded schools receiving accountability grades. Cols (7) and (8): The sample is all non-excluded schools that had not started restructuring prior to accountability. School controls include controls for the average previous year's achievement; the previous year's accountability score (second year only); the percent of students that are black, hispanic, that receive free and reduced price lunch, and that are immigrants; whether the school was in the NYSBP; fixed effects for school size; and five-year average school turnover prior to the institution of accountability. All regressions control for the “RD running variable” (accountability score relative to the threshold) separately by (year)X(schooltype)X(received lower grade), as well as controls for (year)X(schooltype)X(received lower grade) and all of the main effects and lower-order interactions (e.g., year, schooltype, yearXschooltype, etc.). Data come from the 2007-08 and 2008-09 school years in the New York City Department of Education, using the report card grade that was received by the school during fall of the school year. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Appendix Table A3. Covariate Balance

Panel A. Teacher characteristics

	Teacher characteristics												*Turnover*
Dependent Variable:	Math value-added (VA)	ELA VA	ELA predicted VA	Math predicted VA	Female	White	Asian	Black	Hispanic	Years of experience	Has master's degree	Age	Prior turnover
Independent Variable = *School received **lower** grade at the:*	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
Bottom of the grade distribution													
D/F or C/D thresholds (grouped)	0.050	0.012	0.001	0.000	0.009	0.067	-0.006	-0.088	0.028	-0.227	0.002	-0.594	0.008
	[0.10]	[0.08]	[0.00]	[0.00]	[0.02]	[0.06]	[0.01]	[0.05]*	[0.03]	[0.59]	[0.03]	[0.81]	[0.01]
N	306	306	309	309	309	309	309	309	309	309	309	309	309
Dep. Var. Mean	-0.13	-0.09	0.04	0.01	0.81	0.55	0.04	0.27	0.15	9.28	0.41	40.90	0.17
Top of the grade distribution													
B/C or A/B Thresholds	-0.007	-0.003	0.001	0.001	-0.005	-0.046	0.001	0.007	0.039	-0.290	-0.017	-0.069	0.015
	[0.053]	[0.044]	[0.000]	[0.000]*	[0.009]	[0.031]	[0.010]	[0.027]	[0.017]**	[0.310]	[0.017]	[0.457]	[0.008]*
N	927	928	932	932	932	932	932	932	932	932	932	932	932
Dep. Var. Mean	0.00	0.00	0.04	0.01	0.83	0.61	0.05	0.21	0.14	9.72	0.43	40.97	0.157

Panel B. Student and school characteristics

	Student body characteristics									*School characteristics*			
Dependent Variable:	Previous year math scores	Previous year ELA scores	Female	White	Asian	Black	Hispanic	Immigrants	Poverty status	ln(enrollment)	ln(# teachers)	New hires	Teacher bonus program
Independent Variable = *School received lower grade at the:*	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
Bottom of the grade distribution													
D/F or C/D thresholds (grouped)	0.016	-0.030	0.004	0.065	-0.016	-0.185	0.135	0.000	0.011	-0.145	-0.133	-0.194	0.049
	[0.08]	[0.08]	[0.01]	[0.05]	[0.02]	[0.06]***	[0.06]**	[0.00]	[0.05]	[0.12]	[0.11]	[1.12]	[0.10]
N	309	309	309	309	309	309	309	309	309	309	309	309	309
Dep. Var. Mean	-0.25	-0.18	0.49	0.12	0.07	0.42	0.39	0.02	0.71	6.33	3.91	7.78	0.17
Top of the grade distribution													
B/C or A/B Thresholds	-0.051	-0.056	0.003	-0.031	-0.031	0.013	0.050	0.003	0.026	0.088	0.041	0.361	-0.012
	[0.059]	[0.059]	[0.005]	[0.031]	[0.024]	[0.038]	[0.035]	[0.003]	[0.031]	[0.065]	[0.059]	[0.637]	[0.046]
N	932	932	932	932	932	932	932	932	932	932	932	932	932
Dep. Var. Mean	0.03	0.03	0.49	0.15	0.13	0.32	0.39	0.02	0.67	6.43	3.98	6.75	0.16

Notes. This table presents balance tests: regression discontinuity estimates where the outcome variables are teacher (Panel A), student (Panel B, columns (1)-(9)), or school characteristics (Panel B, columns (10)-(13)). Each observation is a school in a given year. All regressions include controls for the "RD running variable" (specifically, a control for (year)X(schooltype)X(received lower grade)) and all of the main effects and lower-order interactions (e.g., year, schooltype, yearXschooltype, etc.), and a dummy for receiving the lower accountability grade at the threshold). Poverty is the percent of students who receive free and reduced price lunch. Prior turnover is the control used in the regressions, which represents 5-year average turnover before the institution of accountability. Standard errors are reported in brackets. Data come from the 2008-09 and 2009-10 school years. All data from the New York City Department of Education. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Appendix Table A4. Joiner and leaver characteristics: Full set of placebo tests

Sample:	Leavers				Joiners			
Independent Var. =	Estimated math VA	Predicted math VA	>=4 years experience	Has a masters	Estimated math VA	Predicted math VA	>=4 years experience	Has a masters
*Received **lower** grade at:*	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A. Both years of accountability system								
Bottom of the grade dist.								
D/F and C/D thresholds	0.107	-0.095	-0.033	-0.073	0.241	-0.087	0.036	0.047
	[0.17]	[0.09]	[0.05]	[0.04]*	[0.26]	[0.09]	[0.04]	[0.03]
N	496	2,118	2,118	2,118	227	2,404	2,404	2,404
Dep. Var. Mean	-0.16	-0.07	0.52	0.34	-0.11	-0.14	0.24	0.18
Top of the grade dist.								
A/B and B/C thresholds	0.016	-0.032	0.012	-0.006	0.324	0.069	-0.039	-0.013
	[0.10]	[0.05]	[0.03]	[0.03]	[0.17]*	[0.06]	[0.02]	[0.02]
N	1,274	5,628	5,628	5,628	521	6,292	6,292	6,292
Dep. Var. Mean	-0.02	-0.02	0.54	0.36	-0.06	-0.09	0.21	0.17
Panel B. First year of accountability system only								
Bottom of the grade dist.								
D/F and C/D thresholds	0.119	0.093	-0.032	-0.118	0.005	0.024	-0.001	0.013
	[0.21]	[0.13]	[0.08]	[0.05]**	[0.31]	[0.10]	[0.05]	[0.05]
N	328	1,392	1,392	1,392	163	1,655	1,655	1,655
Dep. Var. Mean	-0.17	0.00	0.55	0.36	-0.14	-0.11	0.24	0.18
Top of the grade dist.	-0.070	-0.019	0.021	-0.056	-0.036	0.040	-0.030	-0.054
A/B and B/C thresholds	[0.13]	[0.07]	[0.04]	[0.04]	[0.22]	[0.07]	[0.03]	[0.03]**
N	754	3,190	3,190	3,190	334	3,836	3,836	3,836
Dep. Var. Mean	-0.04	-0.03	0.54	0.36	-0.06	-0.11	0.21	0.18

Notes. Table presents regression discontinuity estimates of the effect of school accountability grades on the characteristics of the teachers that leave (leavers) and teachers that are hired by a given school (joiners). Specifically, each observation is a teacher in a given year. The dependent variable is the estimated math value-added (VA) of the teachers in the sample (columns (1) and (5)), the predicted math VA of the teachers (columns (2) and (6)), teachers' experience (columns (3) and (7)), or teachers' education (columns (4) and (8)). The sample is leavers from the school at the end of the year (columns (1)-(4)) and the joiners hired to start the next year (columns (5)-(8)); panel A shows the results using data for both years of the accountability system while panel B displays the results using only the first year. Regressions use a bandwidth of 5 grade points. Standard errors are reported in brackets and clustered at the school level. All regressions control for the "RD running variable" (accountability score relative to the threshold) separately by (year)X(schooltype)X(received lower grade), as well as controls for (year)X(schooltype)X(received lower grade) and all of the main effects and lower-order interactions (e.g., year, schooltype, yearXschooltype, etc.), as well as school controls (which include controls for the average previous year's achievement; previous year's accountability score (second year only); the percent of students that are black, Hispanic, Asian, that receive free and reduced price lunch, and that are immigrants; fixed effects for school size; whether the school was in the NYSBP; and five-year average school turnover prior to the institution of accountability). Data come from the 2007-08 and 2008-09 school years, with the report card grades used being the report card that was received by the school during fall of the school year. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Appendix Table A5. Effect of School Accountability Grades on School Characteristics

Dependent Variable:	Leavers (as % current number teachers)	New hires i.e., joiners (as % current number teachers)	% change in number teachers	% change in enrollment	% change in pupil teacher ratio	Principal turnover
Independent Var. = *School received **lower** grade at the:*	(1)	(2)	(3)	(4)	(5)	(6)
Bottom of the grade distribution						
D/F or C/D thresholds (grouped)	-0.03	-0.01	0.02	0.01	-0.01	0.00
	[0.015]**	[0.015]	[0.018]	[0.029]	[0.032]	[0.079]
N	309	309	309	306	306	290
Dep. Var. Mean	0.14	0.10	-0.03	-0.06	-0.01	0.08
Top of the grade distribution						
B/C or A/B Thresholds (grouped)	0.00	-0.01	-0.01	-0.02	-0.01	0.00
	[0.007]	[0.008]	[0.008]	[0.013]	[0.015]	[0.035]
N	932	932	932	931	931	877
Dep. Var. Mean	0.12	0.09	-0.02	-0.03	-0.01	0.08

Notes. The table presents regression discontinuity estimates of the effect of school accountability grades on school size. Regressions use a bandwidth of 5 grade points. Standard errors are reported in brackets and clustered at the school level. Each observation is a school in a given year. All regressions control for the "RD running variable" (accountability score relative to the threshold) separately by (year)X(schooltype)X(received lower grade), as well as controls for (year)X(schooltype)X(received lower grade) and all of the main effects and lower-order interactions (e.g., year, schooltype, yearXschooltype, etc.) and school controls (which include controls for the average previous year's achievement; previous year's accountability score (second year only); the percent of students that are black, Hispanic, Asian, that receive free and reduced price lunch, and that are immigrants; whether the school was in the NYSBP; fixed effects for school size; and five-year average school turnover prior to the institution of accountability). Columns (1) through (5) estimated weighted by the number of teachers in a school. Data come from the 2008-09 and 2009-10 school years in the New York City Department of Education, using the report card grade that was received by the school during fall of the school year. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Appendix Table A6. Other Robustness Checks for the Regression Discontinuity Turnover Estimates

	Dependent Variable = Teacher Left School						
Independent Var. =	Base sample	Sample excludes phase-out schools	Bounding - includes closing school, assuming 99th percentile turnover	Bounding - includes closing school, assuming 1st percentile turnover	Transfers include mid-year departures	2008 Only	2009 Only
*School received **lower** grade at the:*	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Bottom of the grade distribution							
D/F or C/D thresholds	-0.029	-0.028	-0.029	-0.029	-0.030	-0.027	-0.022
	[0.013]**	[0.013]**	[0.013]**	[0.013]**	[0.014]**	[0.014]**	[0.029]
N	16,897	14,348	16,936	16,936	16,687	11,436	5,461
Top of the grade distribution							
B/C or A/B Thresholds	0.005	0.002	0.005	0.005	0.004	0.010	-0.002
	[0.007]	[0.007]	[0.007]	[0.007]	[0.007]	[0.009]	[0.010]
N	54,672	47,919	54,672	54,672	54,131	28,575	26,097

Notes. The table presents regression discontinuity estimates of the effect of school accountability grades on turnover. Regressions use a bandwidth of 5 grade points. Standard errors are reported in brackets and clustered at the school level. Each observation is a teacher in a given year. Columns (1) and (5)-(7) use the base analysis sample which excludes schools undergoing restructuring; column (2) excludes schools that were in the process of being phased out. Columns (3)-(4) perform a bounding exercise to show whether the one closing school affects the results by including the closing school and assuming either turnover at the 99th percentile of the cross-school distribution (42%) or the 1st percentile (0%). The dependent variable is an indicator that the teacher left the school, either between May of one year and November of the following year (all columns except (5)) or between November of one year and November of the following year (column (5)). Column (6) only includes schools from the 2007-08 school year, and column (7) only includes schools from the 2008-09 school year. All regressions control for the "RD running variable" (accountability score relative to the threshold) separately by (year)X(schooltype)X(received lower grade), as well as controls for (year)X(schooltype)X(received lower grade) and all of the main effects and lower-order interactions (e.g., year, schooltype, yearXschooltype, etc.), school covariates (which include controls for the average previous year's achievement; the previous year's accountability score (second year only); the percent of students that are black, Hispanic, Asian, that receive free and reduced price lunch, and that are immigrants; whether the school was in the NYSBP; fixed effects for school size; and five-year average school turnover prior to the institution of accountability), and teacher covariates (which include fixed effects for teacher experience and age, teacher education level, teacher race, and teacher gender). Data come from the 2007-08 and 2008-09 school years in the New York City Department of Education, using the report card grade that was received by the school during fall of the school year. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Appendix Table A7. Main results estimated with vs. without schools directly on threshold

Description	Turnover		Leaver Value-Added				Joiner Value-added			
Dependent Variable	Teacher Left School		Estimated math VA		Predicted math VA		Estimated math VA		Predicted math VA	
Sample:	Full	Non-restructuring	Full	Non-restructuring	Full	Non-restructuring	Full	Non-restructuring	Full	Non-restructuring
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Bottom of the grade distribution (C/D and										
Base sample	-0.029	-0.030	-0.028	-0.150	0.124	0.021	0.869	0.887	0.164	0.214
	[0.013]**	[0.014]**	[0.21]	[0.24]	[0.10]	[0.12]	[0.34]**	[0.38]**	[0.11]	[0.11]**
N	16,897	14,617	479	385	2,228	1,765	126	113	1,493	1,339
Dep. Var. Mean	0.13	0.12	-0.12	-0.09	-0.07	-0.07	-0.12	-0.12	-0.05	-0.04
Sample Excl. Schools on Threshold	-0.029	-0.031	-0.022	-0.145	0.113	0.004	0.904	0.932	0.181	0.234
	[0.014]**	[0.015]**	[0.21]	[0.25]	[0.10]	[0.12]	[0.39]**	[0.42]**	[0.11]*	[0.11]**
N	16,806	14,526	475	381	2,221	1,758	122	109	1,484	1,330
Dep. Var. Mean	0.13	0.12	-0.11	-0.09	-0.07	-0.07	-0.11	-0.12	-0.05	-0.04
Top of the grade distribution										
Base sample	0.005	0.003	-0.130	-0.145	0.119	0.117	-0.392	-0.374	0.023	0.009
	[0.007]	[0.007]	[0.10]	[0.10]	[0.05]**	[0.05]**	[0.20]*	[0.22]*	[0.07]	[0.07]
N	54,672	47,967	1,194	999	5,916	5,056	379	339	4,251	3,765
Dep. Var. Mean	0.11	0.11	-0.05	-0.05	-0.03	-0.01	-0.09	-0.11	-0.05	-0.04
Sample Excl. Schools on Threshold	0.005	0.003	-0.128	-0.143	0.121	0.118	-0.370	-0.354	0.032	0.019
	[0.007]	[0.007]	[0.10]	[0.10]	[0.05]**	[0.05]**	[0.20]*	[0.22]	[0.07]	[0.07]
N	54,585	47,880	1,189	994	5,900	5,040	377	337	4,239	3,753
Dep. Var. Mean	0.11	0.11	-0.05	-0.05	-0.03	-0.01	-0.09	-0.11	-0.05	-0.04

Notes. Table presents the main estimates from the paper, and then re-estimates those results excluding any schools that fall directly on a grade threshold to address concerns with potential biases that could arise from including those schools. Specifically, columns (1) and (2) replicate the main turnover results (columns (1) and (2) from Table 2), and columns (3) - (10) replicate the main joiner and leaver value-added results (columns (1) - (8) of Table 3), both including and excluding the schools that fall directly on a grade threshold. See table notes from Tables 2 and 3 for the details of the specifications including control variables included. All regressions use a bandwidth of 5 grade points. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Appendix Table A8. Main RD results shown separately for the individual thresholds

Description	Turnover		Leaver Value-Added				Joiner Value-added			
Dependent Variable	Teacher Left School		Estimated math VA		Predicted math VA		Estimated math VA		Predicted math VA	
Sample:	Full	Non-restructuring	Full	Non-restructuring	Full	Non-restructuring	Full	Non-restructuring	Full	Non-restructuring
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Bottom of the grade distribution										
C/D and D/F Pooled	-0.029	-0.030	-0.028	-0.150	0.124	0.021	0.869	0.887	0.164	0.214
	[0.013]**	[0.014]**	[0.21]	[0.24]	[0.10]	[0.12]	[0.34]**	[0.38]**	[0.11]	[0.11]**
N	16,897	14,617	479	385	2,228	1,765	126	113	1,493	1,339
Dep. Var. Mean	0.13	0.12	-0.12	-0.09	-0.07	-0.07	-0.12	-0.12	-0.05	-0.04
D/F Threshold	-0.023	-0.049	-0.252	-0.668	0.318	0.353	2.125	3.382	0.096	0.107
	[0.034]	[0.026]*	[0.33]	[0.39]*	[0.16]**	[0.17]**	[0.92]**	[0.53]***	[0.19]	[0.20]
N	5,194	4,325	185	137	876	630	37	34	470	429
Dep. Var. Mean	0.17	0.15	-0.17	-0.11	-0.05	-0.07	-0.17	-0.13	-0.04	-0.05
C/D Threshold	-0.021	-0.022	0.000	-0.068	0.147	0.026	0.251	0.261	0.162	0.194
	[0.014]	[0.015]	[0.22]	[0.24]	[0.10]	[0.11]	[0.43]	[0.44]	[0.12]	[0.12]
N	13,706	12,100	357	306	1,614	1,359	105	95	1,203	1,080
Dep. Var. Mean	0.12	0.11	-0.08	-0.08	-0.10	-0.07	-0.12	-0.14	-0.05	-0.03
Top of the grade distribution										
A/B and B/C Pooled	0.005	0.003	-0.130	-0.145	0.119	0.117	-0.392	-0.374	0.023	0.009
	[0.007]	[0.007]	[0.10]	[0.10]	[0.05]**	[0.05]**	[0.20]*	[0.22]*	[0.07]	[0.07]
N	54,672	47,967	1,194	999	5,916	5,056	379	339	4,251	3,765
Dep. Var. Mean	0.11	0.11	-0.05	-0.05	-0.03	-0.01	-0.09	-0.11	-0.05	-0.04
B/C Threshold	0.003	-0.001	-0.012	-0.035	0.148	0.150	-0.253	-0.333	-0.021	-0.057
	[0.009]	[0.009]	[0.13]	[0.13]	[0.07]**	[0.07]**	[0.29]	[0.35]	[0.09]	[0.09]
N	29,085	24,411	675	544	3,228	2,676	188	162	2,300	1,993
Dep. Var. Mean	0.11	0.11	-0.11	-0.12	-0.06	-0.03	-0.01	-0.04	-0.07	-0.07
A/B Threshold	0.008	0.004	-0.286	-0.277	0.061	0.058	-0.599	-0.470	0.052	0.073
	[0.010]	[0.010]	[0.14]**	[0.15]*	[0.08]	[0.08]	[0.30]**	[0.28]*	[0.10]	[0.10]
N	25,587	23,556	519	455	2,688	2,380	191	177	1,951	1,772
Dep. Var. Mean	0.11	0.10	0.03	0.02	0.00	0.01	-0.16	-0.17	-0.03	-0.01

Notes. Table presents the main estimates from the paper separately for the individual grade thresholds. Specifically, columns (1) and (2) replicate the main turnover results (columns (1) and (2) from Table 2) separately by grade threshold, and columns (3) - (10) replicate the main joiner and leaver value-added results (columns (1) - (8) of Table 3). See table notes from Tables 2 and 3 for the details of the specifications including control variables included. All regressions use a bandwidth of 5 grade points. ∗ Significant at 10%; ∗∗ significant at 5%; ∗∗∗ significant at 1%.

Appendix Table A9. Other Robustness Checks for the Regression Discontinuity Joiner Quality Estimates

	Dependent Variable = Math Value-Added				
Sample:	Base sample	Sample excludes phase-out schools	2008 Only	2009 Only	Sample excludes joiners from failed schools
Independent Variable = *School received **lower** grade at the:*	(1)	(2)	(3)	(4)	(5)
Bottom of the grade distribution					
C/D and D/F thresholds (grouped)	0.87 [0.34]**	0.88 [0.38]**	0.58 [0.43]	1.14 [0.68]*	1.01 [0.37]***
N	126	112	84	42	122
Top of the grade distribution					
A/B and C/D thresholds (grouped)	-0.39 [0.20]*	-0.37 [0.22]*	-0.25 [0.32]	-0.18 [0.25]	-0.39 [0.20]*
N	379	339	178	201	375
School and teacher covariates	Yes	Yes	Yes	No	Yes

Notes. Table presents regression discontinuity estimates of the effect of school accountability grades on the quality of joiners hired in the subsequent year, using different samples. Each observation is a teacher in a given year, the dependent variable is the math value-added of the teachers in the sample. Regressions use a bandwidth of 5 grade points. Standard errors are reported in brackets and clustered at the school level. Each observation is a teacher in a given year. Cols (1), (3), (4) and (5) use the base analysis sample; col (2) excludes schools that were in the process of being phased out. Column (3) only includes schools from the 2007-08 school year (so teachers that joined at the beginning of the

Appendix Table A10. Descriptive Statistics for the Sample of Schools Not Undergoing Restructuring

	All schools		Not-Restructuring Sample, by Accountability Grade				
	Full sample	Not-Restructuring Sample	A	B	C	D	F
Panel A: Teacher Characteristics							
% Teachers with Master's Degree	0.43	0.44	0.46	0.45	0.43	0.39	0.44
Teacher Experience (years)	9.79	9.81	9.89	10.01	9.69	9.21	9.35
Estimated Math Value-Added	-0.02	-0.01	0.11	0.00	-0.06	-0.11	-0.24
Estimated ELA Value-Added	-0.03	-0.01	0.08	0.00	-0.07	-0.03	-0.16
Predicted Math Value-Added	0.12	0.14	0.17	0.15	0.13	0.06	0.07
Predicted ELA Value-Added	0.09	0.09	0.09	0.08	0.09	0.09	0.01
% Teachers that are:							
Female	0.83	0.84	0.85	0.84	0.84	0.81	0.82
Black	0.20	0.19	0.15	0.18	0.21	0.29	0.22
Non-Hispanic White	0.61	0.63	0.67	0.65	0.61	0.53	0.65
Hispanic	0.14	0.13	0.13	0.12	0.13	0.15	0.09
Asian	0.05	0.04	0.05	0.05	0.04	0.03	0.04
Turnover	0.11	0.11	0.10	0.10	0.11	0.14	0.12
Retirement	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Intra-district transfers	0.04	0.03	0.03	0.03	0.04	0.05	0.05
Exited NYCDOE teacher files	0.07	0.07	0.06	0.06	0.07	0.08	0.07
Sample Size: Teacher-Year Observations (Base Sample)							
All	71,677	62,692	11,263	26,392	18,352	5,728	957
With Math Value-Added Data only	19,092	16,572	2,912	7,081	4,784	1,569	226
Sample Size: Unique Teachers (Base Sample)							
All	50,616	44,303					
With Math Value-Added Data only	13,202	11,499					
Panel B: School Characteristics							
Enrollment	809	783	656	685	667	567	492
% Students that are:							
Black	0.32	0.32	0.28	0.32	0.38	0.46	0.39
Non-Hispanic White	0.15	0.17	0.17	0.17	0.16	0.09	0.18
Hispanic	0.40	0.37	0.37	0.38	0.36	0.40	0.37
Asian	0.12	0.13	0.18	0.13	0.09	0.04	0.06
Immigrants	0.02	0.02	0.01	0.02	0.01	0.01	0.01
Components of Accountability Grades							
Environment Score	7.43	7.58	9.33	8.12	6.97	6.05	6.54
Performance Score	15.11	15.39	18.62	15.86	14.12	11.46	9.65
Progress Score	26.67	26.54	34.31	29.33	21.75	16.02	11.63
Additional Credit	2.18	2.13	3.17	2.31	1.17	0.72	0.24
Overall Score	51.40	51.65	65.45	55.64	44.02	34.26	28.07
Other characteristics							
NY School Bonus Program	0.15	0.13	0.13	0.15	0.14	0.17	0.18
Being phased out[a]	0.01	0.01	0.00	0.00	0.00	0.04	0.05
Not restructuring at baseline[b]	0.87	1.00	1.00	1.00	1.00	1.00	1.00
Sample Size: Schools							
Number of school-year observations	1,243	1,123	207	456	323	115	22
Number of unique schools	847	766					

Notes. Data come from the 2007-08 and 2008-09 school years in the New York City Department of Education. The accountability grade is the school report card grade that was received by the school during fall of the school year. Sample here limited to the sample used for the base turnover analysis: schools within a 5-point bandwidth of one of the grade thresholds. For all columns except the first, sample also limited to schools that were not undergoing restructuring prior to the institution of the accountability system (i.e., that were not in year 2+ of restructuring in the 2007-08 school year).

a. Phase outs proxied for by schools that received accountability grades in one year but not the subsequent year

b. Schools that were not undergoing restructuring prior to the institution of the accountability system (i.e., that were not in year 2+ of

Appendix Table A11. Correlates of teacher mobility

Dep. Var.:	Teacher left school	
Independent Variables:	(1)	(2)
Years of experience	-0.00208***	-0.00125***
	[0.000254]	[0.000227]
Has masters	-0.0180***	-0.0126***
	[0.00324]	[0.00282]
Female	-0.0410***	-0.00860***
	[0.00443]	[0.00322]
Black	0.00424	-0.0424***
	[0.00502]	[0.00390]
Hispanic	-0.0164***	-0.0406***
	[0.00427]	[0.00420]
School Fixed Effects?	No	Yes

Notes. Table presents regressions of teacher mobility on teacher characteristics. Each observation is a teacher in a given year. Data comes from the pre-accountability era (2007 school year). Standard errors clustered at the school level.

* Significant at 10%; ** significant at 5%; *** significant at 1%.

A Value-Added Estimation

To estimate teacher value-added, I follow an approach that has been experimentally validated in the economics of education literature (Kane and Staiger, 2008) and estimate the following regression using the matched student-teacher panel:

$$A_{ijgt} = \alpha + \beta_1 A_{i,j-1,g-1,t-1} + \beta_2 \bar{A}_{-i,j-1,g-1,t-1} + \beta_3 X_i + \tau_t + \tau_g + \tau_j + \eta_{jt} + \varepsilon_{ijgt} \tag{2}$$

where A_{ijgt} is the achievement score (either mathematics or English Language Arts, standardized by year and grade) of student i in the classroom of teacher j in grade g and year t; $A_{i,j-1,g-1,t-1}$ is the student's lagged achievement; $\bar{A}_{-i,j-1,g-1,t-1}$ represents the average previous-year achievement of student i's classmates (to control for peer effects); X_i are student demographics (e.g., gender, ethnicity, eligibility for free-and-reduced-price-lunch); the τ terms represent fixed effects for the year, the grade, and the teacher, respectively; and η_{jt} and ε_{ijgt} represent classroom-level and individual-level error terms, both mean zero and assumed to be independently and identically distributed over time. Following the literature (e.g., Jackson (2012)), in order to have estimates that are comparable across schools, I omit school fixed effects, but the results are quantitatively very similar if I instead estimate using school fixed effects. I only use data from years before the institution of accountability in order to isolate teacher quality from teacher responses to accountability.

I then follow the approach outlined in Kane and Staiger (2008) and Jackson (2009) to create Empirical Bayes (EB) estimates of teacher value-added. Although the estimates obtained by estimating equation 2 directly are consistent (under identifying restrictions), they are not efficient. EB estimates are more efficient, providing the Best Linear Predictor of the random teacher effect in equation 2, which is also the posterior mean with normally distributed errors. Consider the error term in equation 2, $w_{ijgt} \equiv \tau_j + \eta_{jt} + \varepsilon_{ijgt}$. It is the sum of the teacher effect, assumed constant across years, a mean-zero year-specific classroom error, and a mean-zero year-specific student error. To construct EB estimate, I need to estimate the variance of each component. To do this, I first estimate equation 2 using OLS. For the teacher effect, I calculate the mean residual, by teacher, in each year, and use the covariance between these residuals in adjacent years as the estimate of the variance of the teacher effect, $\hat{\sigma_\tau^2} = Cov(\bar{w}_{jgt}, \bar{w}_{jgt-1})$.[47] For the variance of the student effect, I calculate the variance of the student residuals after the classroom mean residual has been removed: $\hat{\sigma_\varepsilon^2} = Var(w_{ijgt} - \bar{w}_{jgt})$. Finally, under the assumption that all three components of the error term are orthogonal to each other, I calculate the variance of the classroom term as the variance of the total error term minus the variance of the teacher and student components: $\hat{\sigma_\eta^2} = Var(w_{ijgt}) - \hat{\sigma_\tau^2} - \hat{\sigma_\varepsilon^2}$.

Next, I compute a raw estimate of a teacher's effect as a weighted average of their classroom residuals ($\bar{w}_{jgt}$), where each classroom is weighted by the inverse of its variance: $\hat{\tau}_j = \sum_{t=1}^{T_j} \bar{w}_{jgt} \frac{(\sigma_\eta^2 + \sigma_\varepsilon^2/N_{jt})^{-1}}{\sum_{t=1}^{T_j}(\sigma_\eta^2 + \sigma_\varepsilon^2/N_{jt})^{-1}}$, where N_{jt} is the number of students in classroom jt and T_j is the total number of classrooms for teacher j.

Finally, I weight this estimate by an estimate of the precision of the teacher's effect to form the empirical Bayes estimate: $\hat{\tau}_j^{EB} = \hat{\tau}_j \frac{\sigma_\tau^2}{\sigma_\tau^2 + [\sum_{t=1}^{T_j}(\sigma_\eta^2 + \sigma_\varepsilon^2/N_{jt})^{-1}]^{-1}}$.

Since the identification of true teacher value-added depends on strong identification assumptions, e.g., that assignment of students to teachers is orthogonal to the student error term ε_{ijgt} in equation 2, recent literature has highlighted the potential biases of value-added measures (e.g., Rothstein (2010)). However, given the RD framework, my identification requirements are less stringent than if I was, say, trying to evaluate teachers based on the estimates. The RD results would only be biased if, conditional on the accountability score, there were differences in the average school-level bias of the value-added estimates that was correlated with the grades. Since the value-added was calculated using pre-period data, this is unlikely. Of greater concern is the comprehensiveness of the value-added estimates: if there are aspects of teacher quality which are not summarized well in teacher value-added measures (which is likely), then my analysis will not incorporate these aspects.

A.1 Calculation of Predicted VA

To compute teacher predicted VA, I follow the approach used in Jackson (2012) and estimate an equation of student achievement (with the inclusion of observable teacher characteristics) on students, using the

[47] This is slightly different from the procedure used by Kane and Staiger (2008) and Jackson (2009), who use the covariance between adjacent classroom-level residuals instead of teacher-level residuals since they both use elementary data only in which the majority of teachers only teach one classroom.

same period of data used above for calculating teacher VA. Specifically, I estimate the following equation:

$$A_{ijgt} = \alpha + \beta_1 A_{i,j-1,g-1,t-1} + \beta_2 \bar{A}_{-i,j-1,g-1,t-1} + \beta_3 X_i + \beta_4 W_{jt} + \beta_5 W_j + \tau_t + \tau_g + \eta_{jt} + \varepsilon_{ijgt} \quad (3)$$

where all variables are defined as before, and the equation now omits the teacher fixed effect τ_j but includes W_{jt} which captures time-varying teacher characteristics (i.e., experience), and W_j which is a vector of observable teacher characteristics (e.g., education and demographics). Using the estimates from equation (3), I predict teacher effectiveness using the observable teacher characteristics. Specifically, the predicted VA for teacher j is $\hat{\beta}_4 W_{jt} + \hat{\beta}_5 W_j$ (i.e., the predicted VA associated with teacher j's experience, education, and demographics). This measure serves as a useful summary statistic for all of the observable teacher characteristics. It is a weighted average of a teacher's observable characteristics, where the weights are determined by the characteristics' relationship with actual student achievement. The advantage of this measure is that it can be calculated for all teachers irrespective of when they enter the data.[48] Unfortunately, I do not have as extensive observable teacher characteristics as some previous papers (e.g., Jackson (2012)) to be able to do the prediction.

B Regression Discontinuity Bandwidth Selection

Since there is no universally agreed-upon method for determining bandwidth for an RD analysis, I follow the standard approach of examining the robustness of the results to different bandwidths.

To select the base bandwidth used for the analyses, I follow the "leave one out" cross-validation procedure of Ludwig and Miller (2007) and Lee and Lemieux (2010) in which I estimate locally linear models at different bandwidths while omitting one observation, calculate the cross-validation criterion as the average squared difference between the predicted and actual values for the omitted observations, and choose the bandwidth that minimizes the cross-validation criterion. The optimal bandwidth using covariates for both the upper and lower thresholds was 5 and so I use that as my base bandwidth for the regressions. When looking at the value-added outcomes and using as my samples either the joiners or the leavers, the optimal bandwidths ranged from 2-7, with the median 4. For consistency, I adopt the

[48] I have test score data from both elementary and middle schools so the measure is defined for all school levels.

turnover bandwidth (5) for my base bandwidth and then show robustness to different bandwidths in the robustness section (Section 6, Tables 9 and 10).

To check robustness, I also calculate a version of the Imbens-Kalyanarman (IK) optimal bandwidth (Imbens and Kalyanaraman, 2012).[49] For the turnover outcomes, these range from 4-6 with a median of 5, and for the value-added, they range from 2-5 with a median of 3; all of these are in the ranges of bandwidths displayed in the robustness tables (Tables 9 and 10).

For the graphs, I show a bandwidth two times wider than the base bandwidth used in the regressions to give a better sense of the regression function.

[49] The IK formula is not developed for the pooled threshold model I use here, where I interact the running variable for indicators for which threshold (school type and year) a given observation is at. I try two modifications to the IK procedure to try to get reasonable estimates within the pooled setting. First, I simply ignore the fact that I am pooling across thresholds, so calculate the IK bandwidth that would be appropriate if there were no interactions with the running variable. Second, I calculate the IK bandwidth separately for each threshold (i.e., for each school type and year). I then calculate the weighted average of the separate bandwidths (weighted by the relative sample sizes), where all are normalized by their sample sizes. Finally, I normalize the averaged bandwidth by the total sample size. In practice, the two methods yield nearly identical results.