How Do School Accountability Reforms Affect Teachers? Evidence from New York City*

Rebecca Dizon-Ross

March 11, 2014

Abstract

A commonly-cited concern with holding schools accountable for student performance is that it could cause good teachers to leave low-performing schools. Using data from New York City, which assigns schools grades based on student achievement, I perform a regression discontinuity analysis and find the opposite effect. At the bottom end of the school grade distribution, a lower accountability grade decreases teacher turnover, especially for high-quality teachers, and increases joining teachers' quality. One potential explanation is that accountability induces performance improvements at lower-graded schools. In contrast, at the top of the grade distribution, where accountability pressures are lower, a lower grade has no turnover effects, but decreases joiner quality.

*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. 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. Contact: Abdul Latif Jameel Poverty Action Lab, MIT, E53-389, 30 Wadsworth St., Cambridge, MA 02142. rdr@mit.edu

1 Introduction

Since the mid-1990s, school accountability systems have become a central focus of education reform efforts 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, make it difficult for low-performing schools to attract and retain good teachers. Evidence from surveys with 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 (Kane et al., 2008; Chetty et al., 2011), this means that accountability systems could have lasting, negative implications for students at low-performing schools. 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, causing more teachers to leave, etc.

The empirical evidence on this question is limited and inconclusive. Several quantitative (Clotfelter et al., 2004; Feng et al., 2010; Li, 2011a) and qualitative (Ladd and Zelli, 2002) papers have suggested that accountability pressures increase teacher and principal turnover at low-performing relative to high-performing schools. Meanwhile, Boyd et al. (2008) find that, when New York introduced state-mandated, high-stakes testing, teacher turnover fell in the grades in which high-stakes exams are administered.[1] The conflicting evidence, as well as methodological limitations of some of the existing work, suggest that further study is needed.

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 based on school performance. The grades were based on continuous performance metrics, with determination of the actual grade based on strict thresholds. This allows the use of 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), the focus here is on the impacts on the teacher labor market.

Using value-added estimates as measures of teacher quality, I find evidence against the

[1] The tests were "high-stakes" because the school-level results were widely disseminated; in the past, tests were primarily used for student evaluation.

hypothesis that accountability makes it harder for low-performing school to attract and retain good teachers. Specifically, using data from the first two years of NYCDOE's accountability system, I find that, at the bottom end of the school grade distribution (i.e., the C/D and D/F thresholds), where the accountability sanctions have more "bite," receipt of a lower school accountability grade early in the school year decreases teacher turnover at the end of the year by over three percentage points. This is a large effect, representing roughly 30% 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 should directly benefit low-graded schools, as turnover has been shown to decrease student achievement (Ronfeldt et al., 2011).

I next examine the sorting implications of accountability grades, and again find that lower accountability grades *help* low-performing schools at the bottom end of the grade distribution. First, lower grades decrease turnover more among high-quality teachers: the turnover of low-quality teachers is in fact unaffected, whereas that of high-quality teachers falls by five percentage points. Second, lower grades increase the quality of joiners.

I examine two main hypotheses to explain the effects: First, that receiving a lower school grade attaches a negative stigma to the teachers at the school, reducing their perceived value to potential employers (the *stigma* hypothesis); and second, that receiving a lower school grade increases the attractiveness of the jobs at the school (the *job desirability* hypothesis). A first potential channel for the (somewhat counterintuitive) *job desirability* hypothesis is school improvement: research has shown that lower-graded schools respond to accountability by improving their performance (e.g., Chiang, 2009), even within the same year the grade was received in New York City (Rockoff and Turner, 2010). Teachers may prefer to teach in schools where achievement is improving, either because they value achievement per se or because they expect that it will lead to higher accountability grades in the future. A second related channel for *job desirability* is that, induced by accountability pressure, school leaders at lower-graded schools work harder to retain their high-quality teachers.

I argue that the *job desirability* hypothesis matches the data better than the *stigma* hypothesis. For example, the fact that lower-graded schools have higher quality joiners than higher-graded schools is more consistent with increased job desirability. The turnover heterogeneity by quality may also support the *job desirability* hypothesis: If one believes that high-quality teachers value performance improvements more than low-quality teachers (e.g., because high quality represents a preference for high achievement) or are differentially the targets of principal retention efforts, then *job desirability* would imply that high-quality teachers would be differentially likely to stay in low-graded schools, which is what we see

here. In contrast, if one thinks that high-quality teachers are less subject to stigma than low-quality teachers (e.g., because they have more qualifications to differentiate themselves on the job market from a stigmatized school), then the *stigma* hypothesis would have the opposite implication. Also suggestive that *stigma* is not driving the results is the fact that lower accountability grades decrease retirements and out-of-district departures.

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 jobs at 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 that grades do not affect teacher turnover or leaver quality, but that there is an asymmetric effect for joiners: lower-graded schools have lower quality joiners than higher-graded schools.

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 the A/B and B/C thresholds) could reflect the fact that accountability pressures are higher at the C/D and D/F thresholds, and so only motivated school improvements there (Rockoff and Turner, 2010).[2] To use these results to assess policymakers' concerns that accountability negatively impacts low-performing schools, one must choose which set of results to focus on. Since the concerns generally focus on the most disadvantaged schools, the effects from the bottom end (C/D and D/F) may be more relevant.

This paper overcomes the two primary challenges in evaluating how accountability affects teachers. The first challenge is identification: accountability reforms 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, which should be similar within small windows. The second challenge is finding good data on teacher quality, and specifically, on teacher's contributions to student learning, or their "value-added." Having a good measure of teacher quality is critical 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. For example, value-added has important predictive power: High value-added improves students' long-run outcomes (Chetty et al., 2011). Unfortunately, value-added estimation has extensive data requirements, and so

[2]Specifically, the explanation is that teachers prefer schools that (1) have improved their environment, performance, or the quality of their teaching positions 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 pressures, and so (1) dominates.

the earlier quantitative papers (Boyd et al. (2008) and Clotfelter et al. (2004)) could only use other teacher characteristics. A long literature has shown that no other characteristics proxy well for value-added (see, e.g., Rivkin et al., 2005; Hanushek et al., 2010).

Feng et al. (2010) is the one previous paper to use value-added data to examine this question. However, the findings presented here stand in stark contrast with their findings. The authors estimate the causal effect of accountability grades on teacher turnover in Florida by exploiting an unexpected change to the school accountability grading system that exogenously "shocked" some schools' grades. They find that receipt of a lower accountability grade increases teacher turnover, with larger effects at the bottom of the grade distribution.

One potential way to resolve the findings in this paper with the causal estimates presented in Feng et al. (2010) is timing. The NYCDOE releases school grades at the beginning of the school year, long before most teachers make turnover decisions, whereas Florida releases grades at the end of the year. If teachers at low-graded schools do not anticipate that their schools will improve, then this could explain the discrepancy, and suggest that delivering grades earlier in the schoolyear provides an easy policy solution to help mitigate accountability's potential negative equity effects.[3] This is, of course, only speculative, as there are many other differences between the two settings.[4]

The 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. It is possible that their results are partially explained by other reforms instituted concurrently with accountability,[5] or by institutional differences, such as the fact that North Carolina linked *teacher-level* incentives with school accountability ratings, whereas the NYCDOE system only used school- and principal-level incentives. The fact that Clotfelter et al. (2004) were unable to use value-added data (and so might be missing important heterogeneity) may also help to explain the qualitatively different findings.

The remainder of the paper proceeds as follows: Section 2 describes the institutional

[3]If teachers in NYCDOE simply thought that "the worst was over" by the summer, i.e., that any negative effects of the accountability grade on their school had already transpired, that could also explain the discrepancies between the findings presented here and those of Feng et al. (2010), but would not explain my finding that turnover actually decreases at lower-graded schools.

[4]Section 2 describes other contextual differences. Another difference between the contexts is that Feng et al. (2010) use a shock to school grades that results from a change to the grading scheme, which could be perceived differently than the introduction of grades into an environment where information about schools' performance is largely unknown/unavailable.

[5]Concurrent reforms include streamlining the process of teacher dismissals and dramatically changing salary structures and tenure requirements. These types of reforms can affect high- and low-performing schools differently, thereby biasing difference-in-differences estimates: for example, the changes to the dismissal process could have increased turnover at low-performing schools if teachers thought dismissals were more likely in those schools.

background. Section 3 provides a conceptual framework. Sections 4 and 5 describe the data and empirical strategy. Section 6 presents the main results, while section 7 discusses potential mechanisms for the results. Section 8 examines robustness and presents specification tests. In section 9, I conclude.

2 Background

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 use the system to improve the performance of low-performing schools.

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 overall 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 a school's score) captures the *level* of standardized test scores. School environment (15% of a school's 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.[6] 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

[6]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.

5th percentiles of pre-additional-credit scores, respectively. Grades are then determined by comparing each school's score to the thresholds.

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).

The sanctions associated with receiving low accountability grades are significant. 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. Schools that are not closed face restricted autonomy, as they must work with their supervisors to develop targeted action plans often involving significant interventions. At the other end of the grade distribution, principals 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.

The NYCDOE Teacher Transfer System

Since 2005, NYCDOE's staffing has been built around the principle of "mutual consent," in which teachers and principals must agree to all teacher placements. This means that the effects estimated in this paper are the effects on equilibrium matches within a market-based system, as opposed to, say, the effects on how administrators make transfer decisions.

Differences Between NYCDOE and Other Contexts

The NYCDOE system had relatively lower paperwork requirements for failing schools than some other accountability systems, such as Florida, where failing schools were required to complete regular, extensive reports (Rouse et al., 2007). As mentioned above, the NYCDOE system also did not link progress reports with teacher-level incentives, whereas North Car-

olina's and Florida's systems included teacher performance bonuses.[7][8] 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.

3 Conceptual Framework

Accountability's effect on teachers will depend on how it affects both teachers' preferences over schools and schools' preferences over teachers. For ease of exposition, I focus here on teachers' preferences. Since I find that accountability pressures decrease turnover, suggesting that the changes are mainly to *voluntary* quits, this is likely to capture the relevant intuition.[9]

Consider the following stylized model describing how teacher i chooses in period t which school to work in, s_{it}, to maximize her utility:

$$\max_{s_{it} \in \mathbf{S}(a_i, g_{s_{it-1}})} U(g_{s_{it}}|a_i) = \max_{s_{it} \in \mathbf{S}(a_i, g_{s_{it-1}})} U(A(g_{s_{it}}), P(g_{s_{it}})|a_i) \tag{1}$$

I assume here that teacher i's utility U from teaching at school s_{it} depends on the teacher's time-invariant quality a_i (proxied in the empirical work with teacher value-added) as well as on the school's end-of-year achievement, $A(g_{s_{it}})$, and the school's prestige, $P(g_{s_{it}})$, both of which depend on the last accountability grade received by the school, $g_{s_{it}}$. Utility is weakly increasing in both achievement and prestige; that is, $\frac{\partial U}{\partial A} \geq 0$ and $\frac{\partial U}{\partial P} \geq 0$. I do not include other school characteristics since, in the empirical analysis, I will be comparing schools on the threshold of grades that were thus ex ante identical.

I assume prestige is increasing in a school's grade $\frac{\partial P}{\partial g} > 0$. However, end-of-year achievement is weakly decreasing in a school's grade ($\frac{\partial A}{\partial g} \leq 0$): many papers have found that lower accountability grades incentivize schools to improve their achievement (Carnoy and Loeb, 2002; Hanushek and Raymond, 2005; Chiang, 2009; Rockoff and Turner, 2010).

The teacher's choice set of schools to work in, $\mathbf{S}$, is assumed to depend on the teacher's ability, a_i, as well as on the grade received by the school at which teacher i was teaching in period $t-1$, $g_{s_{it-1}}$. This captures the fact that teachers coming from schools that received high (low) accountability grades may be seen as more (less) desirable to hire. The choice set also includes the option of leaving the district.

Thus, in period t, teacher i will leave school s_{it-1} (the school that she was teaching at in

[7] 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).

[8] There was a small pilot program for teacher performance pay in the NYCDOE during this time period, but it affected fewer than 20% of schools, and had limited impact, potentially because it was separate from – and viewed as less important than – the accountability system (Li, 2011b).

[9] This simplification still allows schools to make active changes in response to accountability grades, it just does not allow them to fire teachers. There are institutional restrictions that make firing difficult.

period $t-1$) if she could get higher utility from teaching at a different school that is in her choice set, i.e., if the following condition is satisfied:[10]

$$\mathbf{1}\left(U(A(g_{s_{it-1}}), P(g_{s_{it-1}})|a_i) < \max_{s_{it}\in[\mathbf{S}(a_i, g_{s_{it-1}})]} U(A(g_{s_{it}}), P(g_{s_{it}})|a_i)\right) = 1 \tag{2}$$

where $\mathbf{1}$ is the indicator function.

Now, consider two schools: One, s^H, which received a high grade, g^H, coming into period t, and one, s^L, that received a low grade, g^L. Imagine that the schools have identical distributions of teacher ability a_i in the pre-period, $t-1$, which again matches the empirical analysis which compares schools that are ex ante identical but received different accountability grades. In that case, inspection of equation (2) yields the intuitive result that, if turnover decreases at the school which receives a lower accountability grade, it could reflect two mechanisms (which are not mutually exclusive):

1. $U(A(g^L), P(g^L)|a_i) > U(A(g^H), P(g^H)|a_i)$ (the *Job Desirability* Hypothesis):
   Teachers prefer to teach in schools that received lower accountability grades.

2. $\mathbf{S}(a_i, g^L) \subset \mathbf{S}(a_i, g^H)$ (the *Stigma* Hypothesis):
   The quality of a teacher's outside options decrease when the teacher's school receives a lower accountability grade.

Since prestige increases with accountability grades ($P(g^H) > P(g^L)$) while performance decreases ($A(g^H) \leq A(g^L)$), the *job desirability* hypothesis requires that, when comparing a teacher's utility at a lower-graded school to her utility at a higher-graded school, the increase due to higher performance outweighs the decrease due to lower prestige.

One way to try to distinguish between the *stigma* and *job desirability* hypotheses is to see if they have different implications for high-quality and low-quality teachers. Under the *job desirability* hypothesis, the relative turnover by teacher quality would depend on $\frac{\partial}{\partial a_i}\left(U(g^L|a_i) - U(g^H|a_i)\right)$, i.e., on how the gap in utility between teaching in low-graded relative to high-graded schools depends on teacher quality. If we assume that high-quality teachers place a larger value on student performance, that is, that $\frac{\partial^2}{\partial A \partial a}U > 0$ (e.g., because high quality reflects a greater preference for high performance), turnover should fall more for high-quality teachers than low-quality teachers under this hypothesis.[11]

[10] Note that this assumes (1) that teachers are not fired: this is based on the institutional context and the fact that the observed results primarily seem to reflect changes to *voluntary* quits; and (2) that there are no switching costs: the results are robust to adding a switching cost that is fixed across schools.

[11] Note that I am assuming that the high and low quality teachers value prestige equally, or that high-quality teachers value it more but not enough to outweigh their greater preference for performance improvements.

In contrast, under the *stigma* hypothesis, the relative turnover would depend on $\frac{\partial}{\partial a_i}\left(\max_{s_{it}\in[\mathbf{S}(a_i,g^H)]} U(g_{s_{it}}|a_i) - \max_{s_{it}\in[\mathbf{S}(a_i,g^L)]} U(g_{s_{it}}|a_i)\right)$, that is, on whose choice set adjusts more in response to receiving a lower accountability grade. We might expect that the choice sets of low-quality teachers would shrink more than those of high-quality teachers, who may have better individual performance records to differentiate themselves on the job market. In this case, the *stigma* hypothesis would imply the opposite of the *job desirability* hypothesis. I discuss these implications further in Section 7.

Extensions to the Framework

Although the factor that made teachers choose lower-graded schools, A, was labeled above as student performance, A could instead represent any change that accountability pressures induce in lower-graded schools and that increase the teacher's desire to stay at the school. For example, school leaders at lower-graded schools could work harder to retain their teachers by improving their non-financial compensation, or by empowering their teachers. Or, accountability pressures could increase collaboration between teachers as they work to improve. Teachers could even prefer the challenge of the lower grade. It is reasonable to expect that many of these changes would affect higher-quality teachers more (e.g., school leaders likely work harder to retain their high-quality teachers than their low-quality teachers; anecdotally, accountability reforms cause school leaders to differentially empower their high value-added teachers); if so, then the implications are the same as for the *job desirability* hypothesis discussed above.[12]

The empirical analysis also examines how accountability grades affect the quality of teachers joining a school. Generating predictions about this requires moving beyond the very simple framework above and making assumptions about the matching process and schools' preferences over teachers (i.e., on how the choice sets $\mathbf{S}$ are determined). If we assume that true teacher ability is unobservable but that all schools observe a common noisy proxy for ability,[13] then the *job desirability* hypothesis would imply that joiners to low-graded schools would be of higher quality than joiners to high-graded schools, whereas the *stigma* hypothesis would imply the opposite.[14,15]

[12]Principals could also try to push out their low-quality teachers in response to accountability pressure; but this is inconsistent with the result that turnover does not increase at low-graded schools, even for low-quality teachers.

[13]Note that, if ability were perfectly observable, the stigma hypothesis should not play a role (unless schools see accountability grades as informative about a teacher's value conditional on ability).

[14]To distinguish between the hypotheses, I assume a "strong form" of the *stigma* hypothesis in which *job desirability* plays no role, i.e., in which either $A(g^H) = A(g^L)$ or $\frac{\partial}{\partial A}U = 0$, but $P(g^H) > P(g^L)$ and $\frac{\partial}{\partial P}U > 0$. For the *job desirability* hypothesis, I assume that $\frac{\partial}{\partial a_i}\left(U(g^L|a_i) - U(g^H|a_i)\right) \geq 0$, as argued above. I assume a Gale-Shapley-style matching process where the schools propose first (Gale and Shapley, 1962).

[15]This assumes all schools have the same ranking for teachers. Under certain assumptions, the logic goes through if we allow accountability grades to change how schools rank teachers. For example, lower-graded

These joiner predictions implicitly abstract from recruiting; however, accountability pressures could also incentivize low-graded schools to put more effort into recruiting high-quality teachers, a very plausible channel since principals generally devote significant resources to teacher recruitment (Shipps and White, 2009). As long as schools' recruiting efforts for teachers increase with the teacher's rank order, we can think of this as being a variant of the *job desirability* hypothesis.[16]

4 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 breakdown, 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 school 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 school 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 schoolyears. 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 teacher data come from the NYCDOE payroll system and contain teacher experience and salary schedule information, as well as school and grade level identifiers, from the 1999-2000 through the 2009-2010 school years.

I study schools in the first two years that accountability grades were released: the 2007-08 and 2008-09 schoolyears.[17] My sample includes all non-charter elementary, K-8, and middle schools that received accountability grades in the 2007-08 or 2008-09 school years.[18] I exclude all school-year observations where the school closed in the following year (5 observations),

schools may put more emphasis on quality a_i when ranking teachers. If firing is costly enough that low-graded schools do not fire incumbent teachers, this should have the same implications.

[16]That is, we can think of recruiting efforts as increasing the utility of the recruited teachers for the lower-graded school but leaving everyone else's utility unchanged. Since the recruited teachers would be the highest-ranked teachers, who would have on average higher quality, the implications would be the same as those discussed above for the *job desirability* hypothesis. I will not be able to distinguish this variant of the *job desirability* hypothesis in the data from the version described above.

[17]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.

[18]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.

and 6 school-year observations with missing data in the teacher files. To try to remove schools undergoing restructuring, I exclude school-year observations from my base sample that have decreases in staff size or enrollment in the top percentile,[19] but also examine robustness to this exclusion. Descriptive statistics about the schools in the sample are presented in Panel A of Table 1. The sample includes 1,005 unique schools and 1,965 school-year observations.

To estimate teacher value-added, I created a matched-panel of student and teacher data.[20] I use the approach that has been experimentally validated in the economics of education literature (Kane and Staiger, 2008). Appendix A describes the estimation in detail. (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.

Panel B of Table 1 presents descriptive statistics from the sample of teachers teaching in sampled schools in the 2007-08 and 2008-09 school years. The two-year panel contains 61,133 unique teachers and 111,090 teacher-year observations. Roughly 27% of the teachers have math value-added data.[21] Baseline teacher value-added increases slightly with the accountability grade received, with mean value-added at A schools roughly 0.1 standard deviations (of the teacher quality distribution) higher than mean value-added at F schools. Teacher experience and education also both increase with the accountability grade.

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. I also examine robustness to this definition.

Panel A of Table 1 shows that there is 10.7% teacher turnover across the sample period, with turnover increasing across accountability grades from 9.5% at A schools to 14.5% at F schools. Eight percent of the turnover is teacher retirements, 32% is transfers made between

[19]Restructuring is normally a response to long-run trends, not accountability, and so would increase the noise of my estimates (I do not have data on which schools are undergoing restructuring).

[20]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.

[21]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.

teaching positions in the NYCDOE, and 60% reflects departures from NYCDOE.[22]

5 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. (2007), 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:

$$Y_{jt} = \alpha + \beta_g I^g_{jt} + \gamma h(S_{jt}) \times I^g_{jt} \times I^{type}_j \times I_t + \varepsilon_{jt} \tag{3}$$

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^{type}_j, and year, I_t, since the grade thresholds are all specific for school types and years.[23] 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.[24] I also explore robustness to parametric regression functions. All standard errors are clustered at the school level.

The identification assumption is that, conditional on the continuous metric underlying the grade, the grade itself is exogenous. Given 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), this assumption is likely to hold in this context.[25] Unlike in most RD settings, one could also be concerned about ex post "gaming" here: If

[22] 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.

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

[24] 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 Lemiuex (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.

[25] Since all of the score components and the formula for calculating the accountability score are publicly verifiable, the largest potential threat to identification is the fact that the accountability officials who chose the cutoffs were aware of the individual schools and their scores. 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 score 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 cutoffs 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.

administrators took accountability grades into account when selecting schools for closure, this would violate the identification assumption. However, since only one of the schools that was closed during the sample period fell within a 6 point bandwidth of any grade threshold (the largest base bandwidth used in the paper), ex post selection is not driving the results.

As with all RD estimation, the treatment effect under estimation is local to schools adjacent to the grade thresholds, and does not capture any universal effects of accountability.

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.

Density Evidence

The RD design depends crucially on the assumption that there is no manipulation of school scores near the cutoff. Figure 1 plots accountability scores on the X-axis and, on the Y-axis, the number of elementary schools in the 2007/2008 school year that received accountability scores within a 0.5 point bandwidth; the graphs for other schooltypes and years are in the Appendix. It is reassuring that there is not excess density to either side of any of the thresholds.[26] I also perform the density test suggested in McCrary (2008) and do not find evidence of bunching.[27]

6 Results

6.1 Base Turnover Results

I begin with graphical evidence. The left column of Figure 2 plots average residual turnover against accountability scores. Each graph shows a separate grade threshold. To create residual turnover, I regress an indicator for whether a teacher left a school on a vector of covariates.[28] I then group schools according to their accountability scores relative to the grade threshold, 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. Each dot represents 10 schools (it is difficult to see the patterns without some local averaging). For ease of

[26]The density does increase to the right of the D/F threshold, but this does not appear to be excess density but rather a result of the fact that the D/F threshold is drawn at the 5th percentile of the distribution, which is where normal distributions climb precipitously.

[27]I collapse the data to the 0.1 point level, use the "leave one out" procedure to calculate my bandwidth (2 points), and use either the count of schools or count of teachers as the dependent variables, to obtain coefficient estimates for the A-D thresholds of 1.2 (.7), 1.2 (1), .3 (.6), and .6 (.6) using number of schools (standard errors in parentheses) and 69 (54),61(67), 23(46), and 43 (38) using numbers of teachers. None of those coefficients are significant at the 5% level; the coefficient for A using number of schools is significant at the 10% level, but is not robust to the number of teachers specification or to other bandwidths or quadratic or cubic specifications.

[28]The covariates used for creating the residuals are the same used in the regressions and are described in more detail below. Results using the raw outcome variable are very similar and available upon request.

interpretation, I also plot a locally linear regression line in the figures, fitted separately on either side of the grade threshold.

At the C/D and D/F thresholds, there appears to be a small break in turnover at the grade threshold itself: lower-graded schools have locally lower turnover. In contrast, at the A/B and B/C thresholds, there is virtually no discontinuity at the threshold.

To test for the significance of these results, columns (1)-(3) of Table 2 present the regression results, calculated from estimation of equation (3) using an indicator for whether a teacher left the school at the end of the school year as the dependent variable. Each row contains the results from a separate regression. The results presented in the second, third, fifth, and sixth rows use schools within a 6 point bandwidth of a single grading threshold (e.g., A/B, B/C), and control for a linear term in the accountability score, interacted with schooltype, accountability grade, and year. To increase statistical power, the first and fourth rows group the schools from the bottom (C/D and D/F) or top (A/B and B/C) thresholds together; because there are larger increases in the negative accountability pressures placed on schools at the bottom thresholds and smaller marginal increases at the top, the effects should be similar within those groups.[29] All coefficients presented are the coefficient on the indicator that a school received the lower grade at their grade threshold. Column (1) does not control for any covariates. In columns (2) and (3), I add in vectors of school-level and teacher-level covariates.[30] Reassuringly, the estimates are relatively invariant to the addition of covariates, especially in the regressions with more observations (the A/B, B/C, and grouped results).

Consistent with the graphical evidence, the regressions show that, at the bottom end of the grade distribution, lower accountability grades decrease turnover. The grouped threshold result in col. (3) indicates that a lower grade is associated with 3.7 percentage points lower turnover, which is statistically significant at the 1% level. The coefficients for the thresholds estimated separately are similar in magnitude, with point estimates of -4.1 and -2.6 percentage points for the D/F and C/D thresholds, respectively.[31]

A 3.7 percentage point effect is substantial from an economic perspective: It is larger

[29]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 (3), controlling for a linear trend in 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.

[30]School controls include: average student achievement from the previous year; the percent of students that are black, hispanic, that receive free and reduced price lunch, and that are immigrants; fixed effects for school size; and five-year average school turnover before the institution of the accountability system. For the teacher covariates, I use controls that the literature has shown to influence teacher turnover, including fixed effects for teacher experience and age, teacher education level, and teacher gender.

[31]The C/D result is significant at the 5% level as presented; the D/F threshold result is not, but is significant (with similar magnitude) at the 5% level if we use a more parsimonious specification and constrain the effect of the running variable to be the same across years.

than the average turnover gap between D and A schools during my sample period, and is equal to almost 30% of the average school turnover for sample schools in the years preceding accountability.[32] Placing it in the context of the literature on teacher mobility, Hanushek et al. (2004) show that a 10% salary increase decreases teacher turnover for women from 0 to 1.2 percentage points (the vast majority of my sample are female). Thus, even their maximum estimates imply that receiving a lower grade at the C/D or D/F thresholds decreases turnover as much as a 30% increase in salaries would.[33]

This decrease in turnover should provide a direct benefit to low-graded schools, since turnover has been shown to decrease student achievement (Ronfeldt et al., 2011).

In contrast, at the top end of the grade distribution, grades do not affect turnover: none of the estimates are statistically distinguishable from zero, and all are small in magnitude (0.2 - 0.4 percentage points, col. (3)). These are relatively precise zeros: the 95% confidence interval for the grouped threshold result rules out changes of 11% of the mean or 17% of the standard deviation in pre-accountability-era turnover.

I examine the robustness of the turnover results in Section 8.1.

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 teacher outcomes between schools on either side of grade thresholds. The right column in Figure 2 presents these placebo results, plotting average residual turnover in a given year on the accountability score received by a school in the *subsequent* school year. A break in the regression line at the grade thresholds in the year before schools received grades would be concerning. The regression line looks very flat at the A/B and B/C thresholds. There are small breaks at the C/D and D/F thresholds, but they are relatively small, and, at the C/D threshold, go the opposite direction of the actual RD results. Column (4) of Table 2 presents the placebo regression results: reassuringly, none of the placebo coefficients are statistically significantly different from 0.

6.2 Heterogeneity in Turnover by Teacher Quality

The overall turnover results are surprising in the context of the earlier literature (Clotfelter et al., 2004; Feng et al., 2010). To fully understand the implications for low-performing

[32]The average school turnover for the 8 pre-period years for which I have data is 13%, with s.d. of 7.9%.

[33]The literature has also linked student characteristics with teacher mobility, with Scafidi et al. (2007) showing that the percentage of students who are black is the strongest predictor of teacher turnover. Using their estimates, receiving a lower accountability grade decreases turnover as much as decreasing the percentage of black students in a school by one standard deviation.

schools, however, I now look at heterogeneity in turnover by teacher quality: if, say, low-performing teachers are the ones whose turnover falls, then accountability could still make low-performing schools worse off.

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 future 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).[34]

Table 3 presents results from estimating equation (3) separately by different subsamples. Columns (1) and (2) replicate the main findings from Table 2 for the full sample (for comparison) and then for the sample of teachers with math value-added data only. The results in the value-added sample are consistent with the full sample at the C/D threshold; the coefficient at the D/F threshold is actually small and positive, but with a large standard error due to the small sample size.[35]

Columns (3) and (4) of Table 3 present the quality results, showing the results separately for teachers with below-median and above-median math value-added. At the C/D and D/F thresholds, turnover fell more among high-quality teachers at lower-graded schools, with differences of 5.5, 7.9, and 0.9 percentage points at the grouped, D/F, and C/D thresholds, respectively. For the grouped thresholds, high value-added teachers entirely drive the negative turnover result, with the point estimate for below-median value-added teachers near 0 and the effect for above-median value-added teachers 5.5 percentage points more negative and statistically significant at the 10% level.[36]

Thus, accountability helps low-graded schools at the bottom of the grade distribution by not just decreasing turnover, but differentially decreasing turnover of high-quality teachers.

[34]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.

[35]The difference in results at the D/F threshold does not seem to result from different baseline selection into the value-added sample on either side of the grade threshold; instead, it likely results from smaller sample size and statistical noise. There are no statistically significant regression discontinuity effects on whether a teacher has math value-added data: the estimates [standard errors] are -0.012 [0.029] , 0.027 [0.018], -0.012 [0.011], and 0.010 [0.013], for the D/F, C/D, B/C, and A/B thresholds. None of these are statistically significant at the 10% level.

[36]Because the value-added quality data is not available for the majority of the sample, we may also want to look at heterogeneity based on other teacher characteristics. At the D/F and grouped thresholds, the decrease in turnover is much larger for teachers with more than two years of experience than those with less, but the differences are not statistically significant or robust to the C/D threshold or to all measures of experience (e.g., more than 4 years) so it is hard to say much (cols (5)-(8)). The results for teacher education are similar (cols (9)-(10)). Because other characteristics do not proxy well for value-added (Rivkin et al., 2005; Hanushek et al., 2010), we cannot draw firm conclusions, but at least there is no strong evidence that the effects differ in the full sample.

To assess the full effects, I now compare the quality of the teachers that joined the school to that of those who left.

6.3 Quality of Joiners Relative to Leavers

The joiner and leaver quality results are presented in Table 4 and Figures 3 and 4. The left columns of the figures plot the average teacher value-added of leavers (Fig. 3) and joiners (Fig. 4) against school accountability scores. Each dot represents 10 schools. Unfortunately, the density of leaving and new teachers with value-added data at the D/F threshold is too low for these analyses, and so I omit that threshold from the analyses.[37] Table 4 presents the corresponding regressions, calculated by estimating equation (3) using teacher value-added as the dependent variable and including only the leavers or the joiners as the sample. The regressions use a 3 point bandwidth.

Reassuringly, the leaver value-added results are consistent with the heterogeneity results presented in the previous section.

Turning to the joiners, at the bottom end of the grade distribution, Figure 4 suggests and Table 4 (cols. (3) and (4)) confirms that accountability helps lower-graded schools attract higher-quality teachers. Teachers that join D schools have over 1 std. dev. higher value-added than those that join C schools, which is statistically significant at the 5% level and robust to the inclusion of covariates

In contrast, at the top end of the grade distribution (A/B) and (B/C), joiners to lower-graded schools have lower value-added (Fig. 4 and Tbl. 4, cols. (3) and (4)). The effects are large in magnitude, with joiners to B schools of 1.0 std. dev. lower quality than joiners to A schools, and an even larger effect (1.6 std. dev.) at the B/C threshold. The estimates are significant at the 1% and 5% levels, respectively.[38] These results are surprising since there is no effect on turnover or leaver quality at the top of the grade distribution. I discuss potential explanations and mechanisms for this finding in Section 7.

Columns (5) and (6) of Table 4 show regressions where the dependent variable is the value-added of joiners relative to leavers. Consistent with the previous results, the coefficients are positive at the C/D threshold and negative at the A/B and B/C thresholds.

Finally, in columns (7) and (8), I use the year-to-year change in school-average teacher math value-added as the dependent variable. As expected, the overall staff quality rises at the lower-graded school at the C/D threshold, and falls at the A/B and B/C thresholds.

[37] Note that this is not inconsistent with the analysis presented in the previous section of the heterogeneity in turnover based on teacher value-added; that analysis uses the teachers that stay as well as the teachers that leave and so is more robust and appropriate for understanding turnover heterogeneity based on quality. This specification ignores the information on the stayers and so does not perform as well with low school density, but is useful for comparability with the joiner results.

[38] All regressions have at least 47 clusters, and so standard clustering techniques are used.

However, the magnitudes are small and not statistically significant. The magnitudes are small both because overall quality is a stock variable and the flows in a given year are relatively small, and because many joiners are relatively inexperienced teachers who do not have quality data from the pre-period and so do not affect school averages. If the quality patterns are similar in the sample without value-added data, the overall effects would be much larger than indicated by these estimates.

Together, the results imply that, at the bottom end of the grade distribution, lower accountability grades benefit schools by helping them attract and retain high-quality teachers, while at the top end of the grade distribution, they harm schools by decreasing joiner quality.

Quality placebo tests

The right columns of Figures 3 and 4 present placebo results, plotting the average teacher value-added for the leavers and joiners in a given year against the accountability score received by their school in the year *after* the teacher left or joined. It is reassuring that there are no large baseline differences. Columns (9) and (10) of Table 4 show the placebo regression results confirming that there are no baseline differences between the value-added of joiners and leavers at any of the grade thresholds. The robustness of the joiner results is examined in Section 8.2.

6.4 Joiner Results for Other Characteristics

One caveat to the joiner results is that the value-added sample only represents a small fraction of joiners (roughly 10%). Table 5 thus examines whether accountability grades affected other teacher characteristics for which data is available for the full sample. Unfortunately, it has been demonstrated repeatedly in the literature that no observable characteristics proxy well for value-added (Rivkin et al., 2005; Hanushek et al., 2010). However, years of experience is positively correlated, especially in the early years of teaching (Rivkin et al., 2005; Rockoff, 2004), with the second year of teaching associated with the highest increase in quality and the gains to experience generally tapering after 4 years (e.g., Boyd et al., 2008; Kane et al., 2008). Columns (1) - (4) show results using experience as the dependent variable. There are no significant effects at the A/B, B/C, or C/D thresholds, but there is a large effect at the D/F threshold: Joiners to F schools are 45 percentage points more likely to have at least 2 years of experience, and 33 percentage points more likely to have at least 4 years of experience, than joiners to D schools. The effects are significant at the 1% and 5% levels, and robust to using different experience measures (e.g., at least 3 years, 5 years).[39] The result thus are suggestive that there could be increases in value-added for lower-graded schools at the D/F

[39]Results using years of experience are similar but have lower power, potentially because of the nonlinearity of effects in experience.

threshold, like there are at the C/D threshold, although it is obviously not conclusive. The teacher education results presented in Columns (5) and (6) are somewhat different: lower-graded schools hire joiners with lower levels of education at the A/B and C/D thresholds, with no effects at the B/C and D/F thresholds. Since the correlation between education and having a master's degree is not statistically significant in the NYCDOE data, I do not see these results as inconsistent with the value-added results.[40]

Joiner Characteristics Placebo Test

Columns (7) through (9) of Table 5 present placebo regressions using joiner characteristics in the year *before* a school received a given accountability grade. Two of the coefficients for whether a teacher has at least 4 years experience are significant at the 10% level, but none at the 5% level. Since (1) it is only two of the eighteen coefficients presented; (2) the placebo results are not robust to the use of other measures (e.g., at least 5 years experience), and, (3) the results are not significant for the thresholds where we saw results, I do not see these placebo results as cause for concern.

6.5 Summary of Results

At the bottom end of the grade distribution, receiving a lower accountability grade causes teacher turnover to fall at lower-graded schools, with larger decreases for high-quality teachers than low-quality teachers. It also improves the quality of joiners. These results thus provide a much more hopeful story for accountability than many policymakers have feared, implying that, through their labor market effects, accountability systems may actually benefit, not harm, the most disadvantaged schools. In contrast, at the top end of the grade distribution, accountability has no effect on turnover or the quality of leavers, but does decrease the quality of joiners.

In the next section, I examine which mechanisms could explain these findings.

7 Mechanisms

Returning to the framework presented in Section 3, there are two main (non-mutually-exclusive) hypotheses that could explain the finding that receiving a lower accountability grade decreases turnover at the C/D and D/F thresholds:

1. *Job Desirability* Hypothesis: Teachers actively choose to stay in schools that have lower accountability grades.

2. *Stigma* Hypothesis: Lower accountability grades attach a negative stigma to the teachers at the school, thereby decreasing the quality of their outside options.

[40]The link between education and VA is generally tenuous, sometimes even negative (Rivkin et al., 2005).

In the following sections, I provide further motivation for why the *job desirability* hypothesis might hold, evaluate how the evidence presented in Section 6 aligns with the different hypotheses, and then provide further tests.

7.1 Motivating the *Job Desirability* Hypothesis

One potential reason why teachers might choose to stay in lower-graded schools is if lower accountability grades incentivize schools to improve their performance, and teachers like to teach in schools where performance is improving. Indeed, Rockoff and Turner (2010) show us that, in the first year of accountability in the NYCDOE, performance improvements occurred at the D/F threshold even within the same year that the grade was assigned. Columns (1) through (4) of Table 6 replicate Rockoff and Turner (2010)'s results using both the first and second years of accountability data[41] by presenting results from estimation of equation (3) where each observation is a school in a given year, the dependent variable is the school's average standardized English (columns 1 and 2) and math (columns 3 and 4) test scores, and the regressions are estimated with and without controls. Receipt of a lower grade caused F schools to have higher performance than D schools at the threshold. Recall that these improvements happened before teachers made turnover decisions (the official time period for turnover decisions is May through August), and so could have influenced turnover. Although we do not see the same test score increases at the C/D schools in Table 6 (the coefficient estimates are in fact negative but not significant at the 5% level), it is possible that similar cultural or instructional shifts happened at these schools but, because those schools faced lower accountability pressures, the changes were more minor and so did not cause short-run test score improvements (a theory consistent with the smaller turnover effect at the C/D threshold relative to the D/F).

As discussed in Section 3, there are many other ways besides performance improvements that accountability pressures could induce schools to become more attractive to teachers and thus drive the *job desirability* hypothesis (e.g., school leaders could improve teachers' non-financial compensation). I will not be able to separately identify these in the data. There are also other potential reasons besides these types of "improvement stories", but I see these as more plausible and so focus on them below.[42]

[41] The second year of data was not yet available when they wrote their paper.

[42] A first alternative is 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 in fact dislike being substitutes. If we think this explanation is less likely for low-quality teachers, this would also not be consistent with the heterogeneity results in Section 6.2. A second alternative explanation is class size: Since accountability allows students to

7.2 Alignment of Previous Evidence with Hypotheses

I now summarize how the evidence presented in Section 6 aligns with the different hypotheses. The finding that turnover decreased at lower-graded schools at the bottom but not the top of the grade distribution could be easily explained by the *job desirability* hypothesis if the change in accountability pressures when crossing a grade threshold – and thus the pressure-induced increases in job desirability– were larger at the bottom end. This seems plausible based on the institutional guidelines and the Table 6 results. In contrast, although certainly possible, it is not clear why one would expect ex ante that stigma would change more when moving across grade thresholds at the bottom end of the grade distribution than at the top end, as the *stigma* hypothesis would require.

The *job desirability* hypothesis is also consistent with the second finding – that turnover fell more among high-quality teachers than low-quality teachers at the bottom end of the grade distribution – since high-quality teachers may place higher value on school performance improvements than low-quality teachers, benefit more from the enfranchisement of high value-added teachers, and/or be the focus of principals' increased retention efforts. In contrast, if we believe that high-quality teachers can better differentiate themselves from a stigmatized school on the job market, the *stigma* hypothesis is less consistent with this finding.

Third, the *job desirability* hypothesis would imply that, at the grade thresholds where we saw turnover effects, joiners to lower-graded schools would be of higher quality, which is exactly what we see. This could either reflect the schools becoming more attractive or school leaders working harder to recruit. The *stigma* hypothesis would have implied the opposite.

Either theory is consistent with the fact that, at the top end of the grade distribution, where we saw no turnover effects, joiner quality is lower. Under the *job desirability* hypothesis, this would imply that the fall in prestige was more salient to individuals seeking jobs – who may start their job search by looking up school grades– than incumbents – who know much more about a school than its accountability grade. (The *job desirability* hypothesis is not that lower grades do not come with stigma but rather that that stigma is not the primary factor causing a fall in turnover at lower-graded schools.)

Thus, the evidence presented so far is more aligned with the *job desirability* hypothesis, although the arguments are somewhat speculative. The next section provides additional

transfer out of F schools, if teachers prefer smaller classes, they could stay in lower-graded schools. However, columns (3) and (4) of Appendix Table 2 present regressions where the dependent variable is the percent change in school enrollment after receiving an accountability grade. None of the coefficients at the C/D or D/F thresholds are significant at even the 10% level (the B/C threshold coefficient is significant at the 10% level, but would not explain the effect since we do not see decreases in turnover at the B/C threshold). The estimate at the grouped C/D and D/F thresholds shows that enrollment at lower-graded schools decreased by 2%; it is unlikely that such a small drop could cause such a large decrease in turnover.

suggestive evidence supporting that hypothesis.

7.3 Additional Tests of the Hypotheses

Turnover: Destinations

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 7 breaks down 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.[43] The turnover result is driven almost entirely by fewer teachers leaving NYCDOE (col. (4)), which accounts for roughly 80% of the decrease in turnover at lower-graded schools. This is larger than the share of overall turnover driven by departures from the NYCDOE (60%). In contrast, within-district transfers (col. (3)) represent a smaller percentage of the effect than of overall turnover. (Note that in neither case can we reject equality.) In addition, we see a statistically significant effect on retirements at the D/F threshold, which we would not expect to be affected by stigma. Thus, the table also supports the *job desirability* hypothesis.

Relationship Between Performance and Turnover

The argument that performance improvements may underlie the *job desirability* hypothesis depends on the assumption that teachers prefer to teach in schools that have improved their performance in response to accountability. We can evaluate how plausible this assumption is by testing whether turnover falls at schools that improve their achievement. (Note that we do not want to look at RD estimates of the effects of grades: the hypothesis is not that the effect of school improvements differs by grade, but rather that grades cause school improvements which in turn cause lower turnover.) Columns (1)-(3) of Table 8 present results from regressions of teacher turnover at the end of the school year on the school's average student achievement in the same year. Across the grade distribution, we see that turnover is lower when schools have higher achievement (conditional on prior achievement), providing suggestive evidence that performance improvements could be one mechanism through which lower accountability grades decrease turnover.[44]

[43] 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.

[44] Each row shows results for schools within a small bandwidth of different grade thresholds (the same bandwidth used in the RD specifications) in order to keep the sample comparable to the RD estimates. The results are quantitatively similar if the estimation is performed separately on either side of the grade threshold, e.g., including F schools only instead of both F and D schools near the D/F threshold, standard

A second test is to look at whether schools that improve attract higher-quality teachers. Columns (4)-(6) of Table 8 show regressions of joiner value-added on school achievement. Schools that improve more have higher joiner quality.

Note that this evidence does not help distinguish between the potential channels underlying the *job desirability* hypothesis (e.g., performance improvements vs. recruitment efforts) because they are likely highly correlated.

8 Robustness of the RD Results

8.1 Robustness of the Turnover Results

Table 9 shows that the turnover findings are not due to the particular RD specification used. Columns (1) through (10) present the results using linear specifications with a range of bandwidths (specifically, 50% and 200% of the base bandwidth, as well as the base bandwidth +/-1), and the results are similar. Columns (11) through (14) 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), especially the A/B, B/C, and grouped C/D D/F results. Columns (15) and (16) show robustness to controlling linearly for all of the components of the accountability score separately instead of the composite score.[45]

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.[46] Reassuringly, the test statistic is not rejected in any locally linear specification. It is rejected in a few quadratic and cubic specifications; since this test can be used to evaluate the appropriate control function, this suggests that the linear specification is correct here.

Columns (1) and (2) of Appendix Table 1 demonstrate robustness to the sample selection criteria, showing that the results are similar, if statistically weaker, when one includes outliers in the sample, while columns (3) and (4) show that the results are robust to counting midyear departures as turnover.

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

errors are just larger due to smaller sample size. Column (1) runs the regressions with no covariates; column (2) adds in school and teacher covariates– critically, including a control for previous-year achievement– and column (3) adds accountability score controls.

[45]This is the approach adopted by Rockoff and Turner (2010), but I do not do this in my base specifications since I use smaller bandwidths and so a more parsimonious specification is preferable.

[46]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. Note that this test can also be seen as a test for whether the regression function is well approximated by the linear function within the bandwidth.

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.[47] Columns (5)-(6) of the table show that, reassuringly, the results are qualitatively similar, but noisier and less robust (as would be expected given the smaller sample sizes), when one estimates equation 3 separately for the 2007-08 and 2008-09 school years.

Appendix Table 2 demonstrates that the observed turnover effects do not result mechanically from a change in staff size brought on by accountability grades.[48]

8.2 Robustness of the Joiner Quality Results

Table 10 examines the robustness of the joiner quality results. Columns (1) through (10) show that the B/C and C/D estimates are relatively stable across different bandwidths, with the C/D effect concentrated close to the threshold as it fades a little with higher bandwidths. The A/B result is somewhat more sensitive to bandwidth. Columns (11) through (14) show the results using quadratic or cubic control terms. The B/C result is also robust to these specifications; the A/B and C/D results maintain their signs, but lose some of their magnitude and statistical significance. Columns (15) and (16) show that the qualitative findings are robust to using the components of the accountability score instead of the composite score. Columns (17) - (20) demonstrate robustness to the value-added functional form, as the results are qualitatively similar using an indicator for a teacher being above median value-added or Empirical Bayes estimates as the dependent variables. I also perform the same specification test described in section 8.1 (Lee and Lemieux, 2010). The p-value for the test statistic is above 0.10 for all specifications. Finally, if one were concerned that the effects reflect differential selection into the value-added sample on either side of the thresholds, I also estimate whether there is an RD effect on whether a new teacher hired has value-added data and find no significant effects at any of the grade thresholds.

[47] 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.

[48] Specifically, columns (1) and (2) of Appendix Table 2 present regressions of equation (3) where the dependent variable is the percent change in the number of teachers at a given school after receiving an accountability grade (each observation is a school-year). All coefficients are small in magnitude, none are statistically significant, and some have a positive sign, which means the mechanical effect of changing staff size on turnover would go the opposite direction from the observed turnover effects.

9 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), accountability positively impacts lower-graded schools by decreasing turnover, especially among high-quality teachers, and by increasing the quality of joiners. A plausible explanation is that teachers actively choose to stay in the lower-graded schools because job desirability increases at those schools, perhaps because academic performance has improved, or because school leaders put more effort into attracting high-quality teachers. In contrast, at the top end of the grade distribution (the A/B and B/C thresholds), where the accountability pressures are relatively low, I find that receiving a lower accountability grade does not change the quantity or quality of the leavers, but does decrease joiner quality. This could imply that the nominal accountability grade matters more to joiners than leavers because they have less other information about school quality, and that, all else equal, teachers prefer schools with higher nominal grades.

This paper provides direct evidence against one of the major concerns with accountability systems: that they would have negative equity effects through the teacher labor market. Instead, the results imply that accountability's labor market effects promote equity, and, over time, could help narrow the distribution of school performance and accountability grades as better teachers move to schools that were lower-performing at baseline. (Extrapolating over time is of course speculative.) A related, common concern with accountability systems is that they could cause teachers to leave the profession as "too much pressure [would] lead to dissatisfaction [and] exit" (Mintrop and Trujillo, 2005). Although we should be cautious in extrapolating from partial to general equilibrium, the results presented here suggest that the concern may be unfounded. That is, the fact that both turnover and retirements fell at the schools that faced the highest accountability pressures suggests that, if anything, the general equilibrium impact of accountability would be to decrease overall quit rates of teachers from teaching, thus potentially increasing average teacher experience and quality.

This paper thus presents a hopeful message for accountability. This stands in contrast to much of the earlier literature, especially Feng et al. (2010) and Clotfelter et al. (2004). 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 (e.g., the timing of grade release, whether the incentives are targeted at the teacher level). Better understanding of these features would enable policymakers to continue to design accountability systems that 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.

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

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

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.

Feng, L., D. N. Figlio, and T. Sass (2010, June). School accountability and teacher mobility. National Bureau of Economic Research Working Paper No. 16070.

Gale, D. and L. S. Shapley (1962). College admissions and the stability of marriage. *American Mathematical Monthly*, 9–15.

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

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

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. Lemiuex (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics 142*(2), 611–650.

Jackson, C. and E. Bruegmann (2009). Teaching students and teaching each other: The importance of peer learning for teachers. Technical report, National Bureau of Economic Research.

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.

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 Kappan 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). Are teacher-level value-added estimates biased? An experimental validation of non-experimental estimates. Unpublished Manuscript, Harvard University.

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

Ladd, H. F. and A. Zelli (2002). School-based accountability in north carolina: The responses of school principals. *Education Administration Quarterly 38*, 494–529.

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

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

Li, J. (2011b). What New York City's experiment with schoolwide performance bonuses tells us about pay for performance. Research Brief. *RAND Corporation*.

Ludwig, J. and D. L. Miller (2005). Does Head Start improve children's life chances? Evidence from a regression discontinuity design. Technical report, National Bureau of Economic Research.

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.

Mintrop, H. and T. Trujillo (2005). Corrective action in low performing schools: Lessons for NCLB implementation from first-generation accountability systems. vol. 13 no. 48. *education policy analysis archives 13*, 48.

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*.

Ronfeldt, M., H. Lankford, S. Loeb, and J. Wyckoff (2011). How teacher turnover harms student achievement. Technical report, National Bureau of Economic Research.

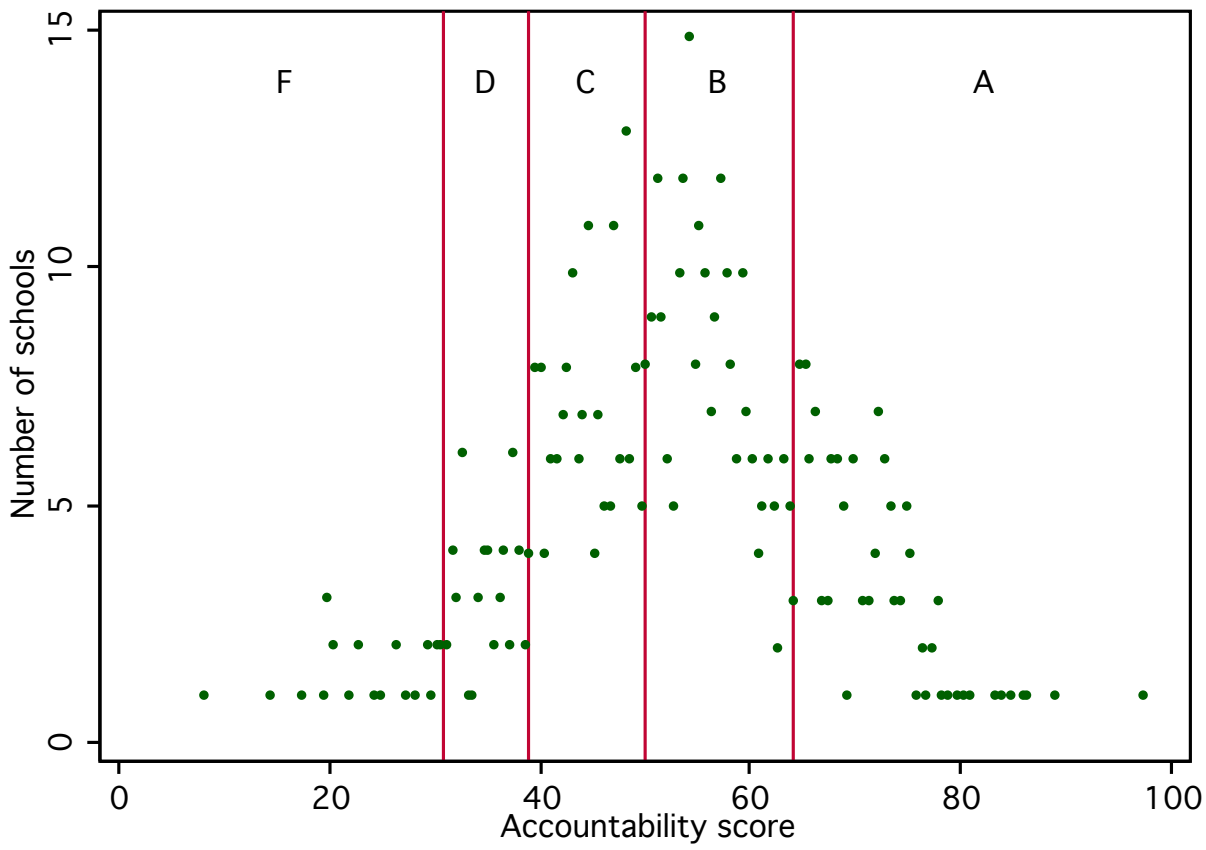
Rothstein, J. (2010, February). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics 125*(2), 175–215.

Rouse, C. E., J. Hannaway, D. Goldhaber, and D. N. Figlio (2007). Feeling the Florida heat? How low-performing schools respond to voucher and accountability pressure. National Bureau of Economic Research Working Paper No. 13681.

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

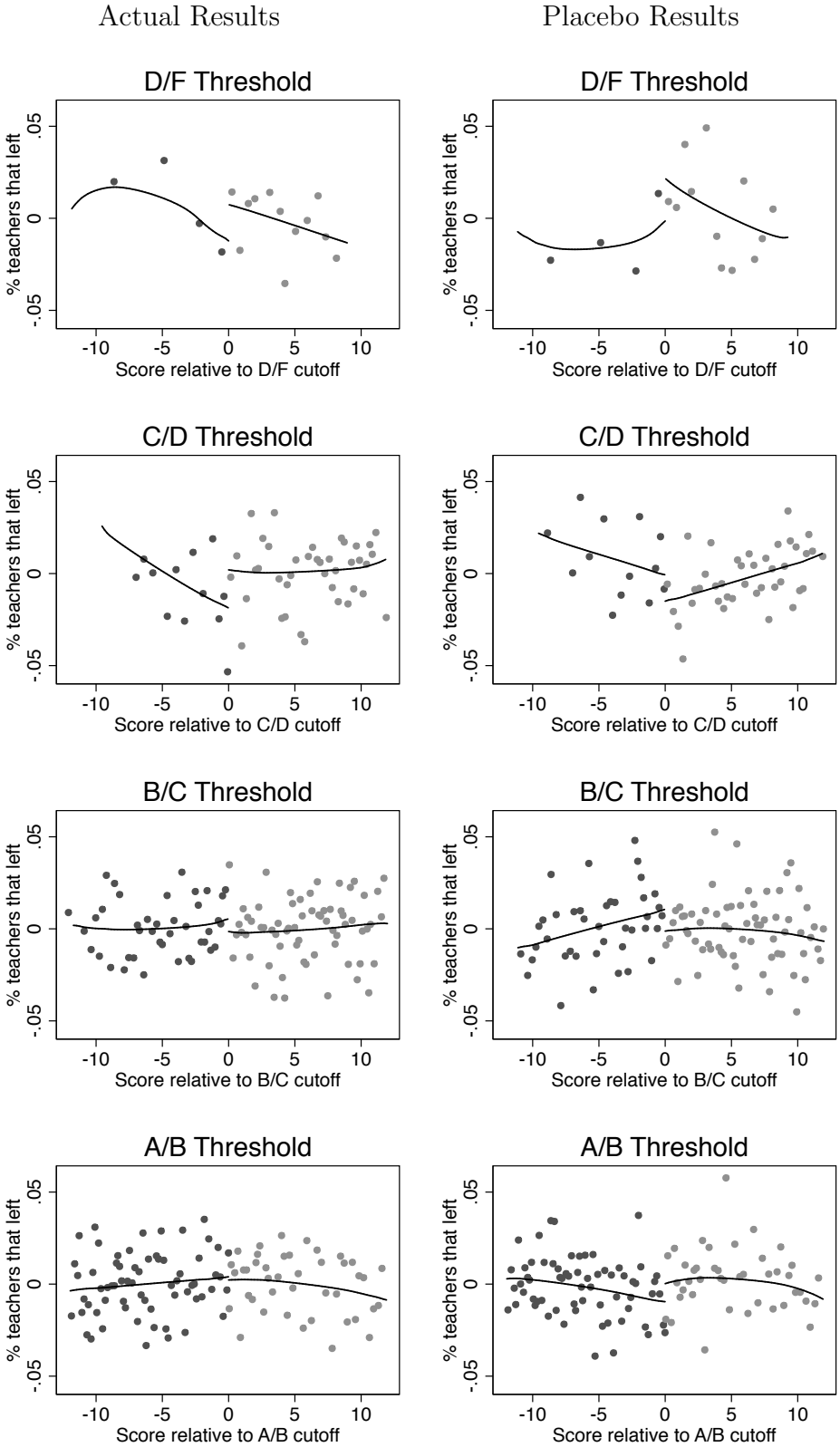
Shipps, D. and M. White (2009). A new politics of the principalship? Accountability-driven change in New York City. *Peabody Journal of Education 84*(3), 350–373.

Figure 1: Density of Schools Near the Accountability Grade Thresholds



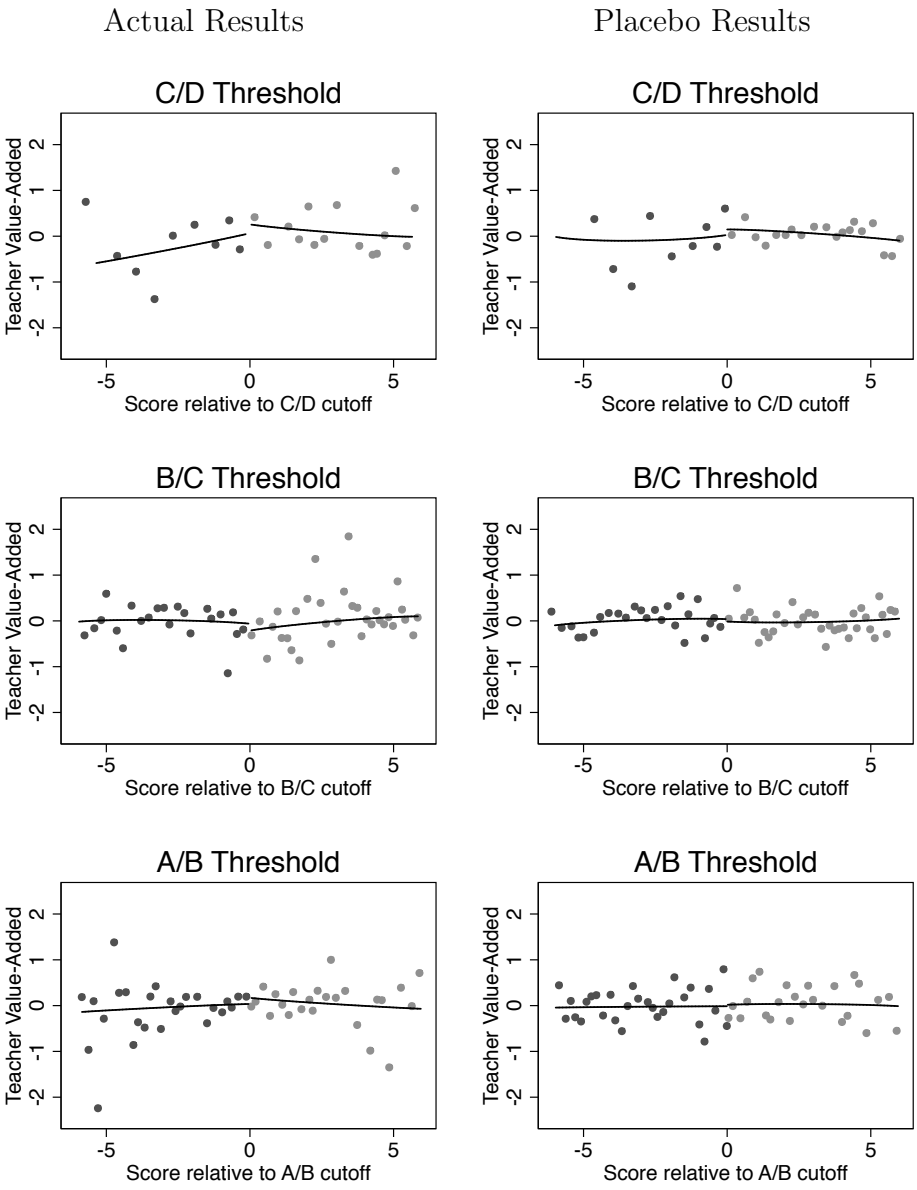
Notes. The figure plots the number of elementary schools in 2008 with a given accountability score (specifically, the y-axis shows the number of schools within a 0.5 point bandwidth of the accountability score displayed on the X-axis). The red lines show the 4 grade thresholds (A/B, B/C, C/D, and D/F). Evidence of heaping directly adjacent to the grade thresholds line would be a violation of the regression discontinuity identification assumptions.

Figure 2: Residual Turnover, by Accountability Score



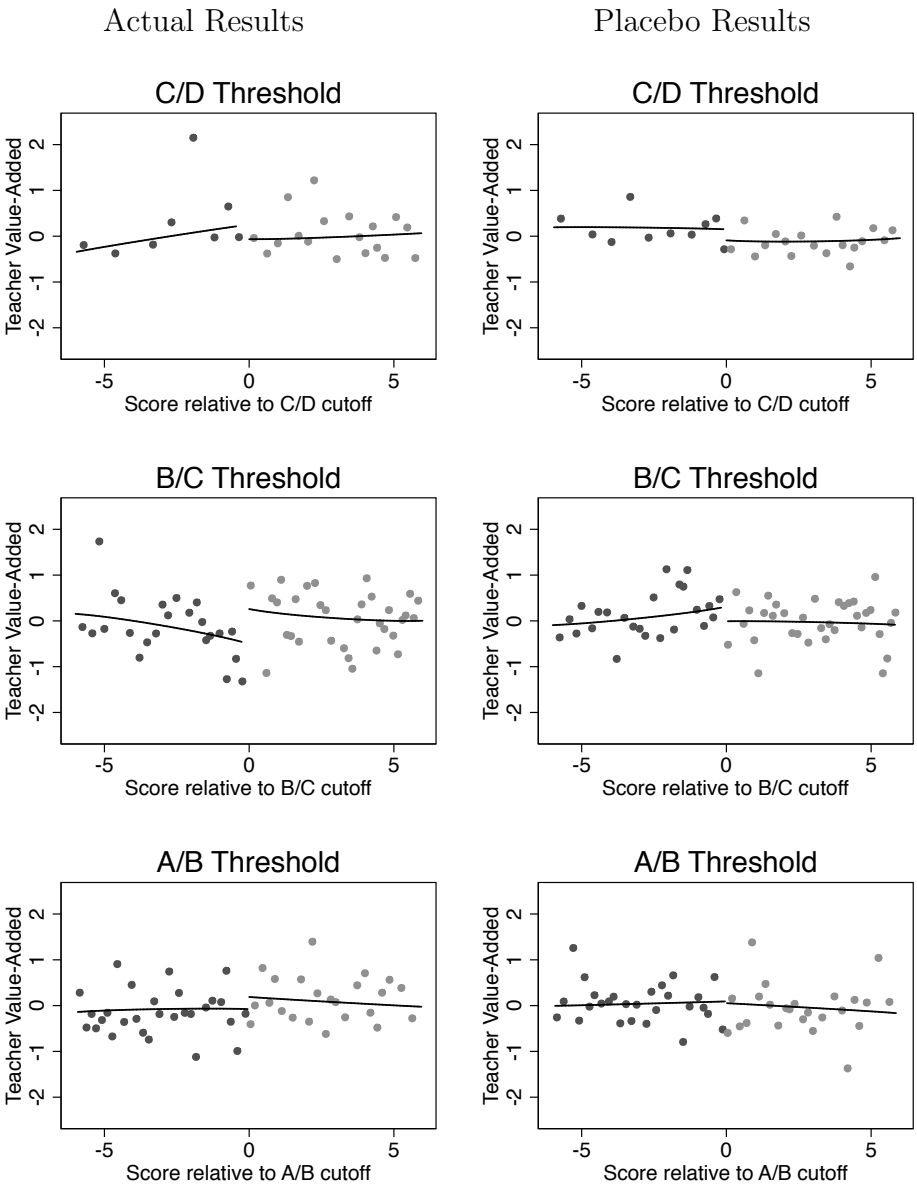
Notes. The left column plots the actual turnover results, plotting average residual turnover in the summer after a school received an accountability grade on the school's accountability score relative to the grade threshold (so the grade threshold is always displayed at 0). Each dot represents 10 schools. The right panel has placebo turnover results: there, the y-axes show residual turnover in the year *before* a school received an accountability grade. 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).

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



Notes. The left column plots the actual leaver quality results. The x-axes show schools' 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). Each dot represents 10 schools. 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.

Figure 4: Average Math Value-Added of Joiners, by Accountability Score



Notes. The left panel plots the actual joiner quality results. The x-axes show schools' 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 joiners (i.e., of the teachers who joined schools in the summer after their schools received the accountability score and grade). Each dot represents 10 schools. 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.

Table 1. Descriptive Statistics by Accountability Grade

	Accountability Grade					
	A	B	C	D	F	All Schools
Panel A: Teacher Characteristics						
Teacher Math Value-Added	0.07	-0.03	-0.02	-0.13	-0.01	0.00
Teacher ELA Value-Added	0.07	-0.01	-0.07	-0.03	-0.02	0.00
% Teachers with Master's Degree	45%	44%	43%	39%	40%	44%
Teacher Experience (years)	9.9	9.9	9.7	9.2	9.3	9.8
% Teachers that are:						
Female	85%	83%	82%	80%	81%	83%
Black	15%	20%	23%	29%	29%	20%
Non-Hispanic White	65%	62%	59%	51%	55%	61%
Hispanic	14%	14%	14%	16%	12%	14%
Asian	5%	5%	4%	4%	3%	5%
Turnover	0.10	0.11	0.11	0.13	0.15	0.11
Retirement	0.01	0.01	0.01	0.01	0.01	0.01
Intra-district transfers	0.03	0.03	0.04	0.04	0.06	0.03
Exited NYCDOE teacher files	0.06	0.06	0.07	0.07	0.07	0.06
Sample Size: Teacher-Year Observations (Base Sample)						
All	32,733	45,784	23,847	6,611	2,115	111,090
With Math Value-Added Data only	8,697	12,234	6,323	1,829	502	29,585
Sample Size: Unique Teachers (Base Sample)						
All						61,133
With Math Value-Added Data only						15,625
Sample Size: Teacher-Year Observations (Joiner Sample)						
All	2,378	3,325	1,828	579	203	8,272
With Math Value-Added Data only	209	271	151	40	22	690
Panel B: School Characteristics						
Enrollment	780	833	802	710	553	798
% Students that are:						
Black	26%	33%	38%	46%	46%	31%
Non-Hispanic White	15%	15%	15%	9%	11%	15%
Hispanic	41%	40%	38%	41%	38%	41%
Asian	17%	12%	9%	4%	4%	13%
Free and Reduced Price Lunch Recipients	2%	2%	2%	1%	1%	2%
Components of Accountability Grades						
Environment Score	9.77	8.11	6.94	5.93	5.63	7.96
Performance Score	18.88	15.75	13.93	11.43	10.54	15.88
Progress Score	39.12	29.63	21.98	16.31	6.62	29.48
Additional Credit	4.31	2.36	1.19	0.67	0.32	2.77
Overall Score	72.08	55.87	44.05	34.35	23.11	56.1
Sample Size: Schools						
Number of school-year observations	599	781	410	126	49	1,965
Number of unique schools						1,005

Notes: Data comes 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.

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

	Dependent Var=1{Teacher Left School}			
	Current Year (Actual Results)			Previous Year (Placebo)
*Independent Var.= School received **lower** grade at the:*	(1)	(2)	(3)	(4)
Bottom of the grade distribution				
D/F or C/D thresholds (grouped)	-0.027	-0.040	-0.037	0.011
	[0.015]*	[0.012]***	[0.012]***	[0.015]
N	17,932	17,932	17,932	17,810
D/F Threshold	-0.078	-0.050	-0.041	0.014
	[0.026]***	[0.029]*	[0.028]	[0.034]
N	5,392	5,392	5,392	5,405
C/D Threshold	-0.009	-0.026	-0.026	0.020
	[0.015]	[0.012]**	[0.012]**	[0.014]
N	15,275	15,275	15,275	15,159
Top of the grade distribution				
B/C or A/B Thresholds (grouped)	0.006	0.003	0.002	0.006
	[0.007]	[0.006]	[0.006]	[0.007]
N	63,970	63,968	63,968	63,739
B/C Threshold	0.002	0.003	0.003	0.014
	[0.010]	[0.008]	[0.008]	[0.009]
N	34,386	34,386	34,386	34,392
A/B Threshold	0.015	0.005	0.004	0.000
	[0.010]	[0.009]	[0.009]	[0.009]
N	30,031	30,029	30,029	29,787
Dependent Variable Mean	0.107	0.107	0.107	0.121
Control for score*year*schooltype* accountability grade?	✓	✓	✓	✓
School covariates?		✓	✓	✓
Teacher covariates?			✓	✓

Notes. Table presents regression discontinuity estimates of the effect of school accountability grades on teacher turnover. In the current year (actual results) regressions, the dependent variable is an indicator for whether a teacher stopped teaching at the school in the summer after the accountability grade was received; in the previous year (placebo) regression, the dependent variable is an indicator for whether a teacher stopped teaching at the school in the summer before 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 6 grade points. Standard errors are reported in brackets and clustered at the school level. School covariates include controls for the average previous year's achievement; the percent of students that are black, hispanic, that receive free and reduced price lunch, and that are immigrants; 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, and teacher gender. Data come from the 2008-09 and 2009-10 school years for the actual regressions and the 2007-08 and 2008-09 school years 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. Heterogeneity in the Regression Discontinuity Turnover Estimates by Teacher Characteristics

	Dependent Variable=1{Teacher Left School}									
			Math Value-Added		Years of Experience		Years of Experience		Education	
Independent Var. =	Full Sample	Teachers with Math Value-Added Data	Below-Median	Above-Median	Less than 2	2 or more	Less than 4	4 or more	No Masters	Masters and above
*School 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 thresholds (grouped)	-0.037	-0.017	0.008	-0.047	-0.015	-0.041	-0.038	-0.036	-0.034	-0.040
	[0.012]***	[0.019]	[0.027]	[0.027]*	[0.039]	[0.012]***	[0.029]	[0.010]***	[0.018]*	[0.014]***
N	17,932	4,789	2,721	2,068	2,988	14,944	5,508	12,424	10,482	7,450
D/F Threshold	-0.041	0.018	0.057	-0.022	0.020	-0.053	-0.030	-0.046	-0.006	-0.084
	[0.028]	[0.035]	[0.054]	[0.055]	[0.090]	[0.028]*	[0.063]	[0.024]*	[0.036]	[0.038]**
N	5,392	1,416	843	573	1,031	4,361	1,792	3,600	3,268	2,124
C/D Threshold	-0.026	-0.033	-0.032	-0.041	-0.055	-0.022	-0.033	-0.027	-0.025	-0.024
	[0.012]**	[0.020]	[0.026]	[0.028]	[0.037]	[0.012]*	[0.030]	[0.011]**	[0.018]	[0.014]*
N	15,275	4,136	2,339	1,797	2,444	12,831	4,581	10,694	8,864	6,411
Top of the grade distribution										
B/C or A/B Thresholds (grouped)	0.002	0.005	0.008	0.001	-0.018	0.006	-0.011	0.008	0.005	-0.001
	[0.006]	[0.009]	[0.012]	[0.012]	[0.021]	[0.006]	[0.014]	[0.006]	[0.008]	[0.007]
N	63,968	17,026	8,822	8,204	9,147	54,821	18,309	45,659	35,821	28,147
B/C Threshold	0.003	-0.003	0.010	-0.013	-0.050	0.011	-0.019	0.010	-0.002	0.006
	[0.008]	[0.012]	[0.016]	[0.016]	[0.028]*	[0.008]	[0.019]	[0.008]	[0.011]	[0.009]
N	34,386	9,262	4,899	4,363	5,080	29,306	9,937	24,449	19,450	14,936
A/B Threshold	0.004	0.015	0.017	0.012	0.030	0.000	0.003	0.003	0.015	-0.011
	[0.009]	[0.013]	[0.018]	[0.019]	[0.029]	[0.009]	[0.020]	[0.009]	[0.012]	[0.012]
N	30,029	7,903	3,981	3,922	4,121	25,908	8,486	21,543	16,608	13,421
Dependent Variable Mean	0.107	0.083	0.085	0.080	0.200	0.092	0.175	0.080	0.123	0.088
Control for score*year* schooltype*grade?	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
School and teacher covariates?	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Notes. Table presents regression discontinuity estimates of the effect of school accountability grades on teacher turnover, estimated separately for different subsamples to show the heterogeneity. 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 varies by column (as indicated in the column heading), and each observation represents one teacher in a given year. Regressions use a bandwidth of 6 grade points. Standard errors are reported in brackets and clustered at the school level. School controls include controls for the average previous year's achievement; the percent of students that are black, hispanic, that receive free and reduced price lunch, and that are immigrants; 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, and teacher gender. 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 4. Regression Discontinuity Estimates of the Effect of School Accountability Grades on the Quality of Leavers and Joiners

	Dependent Variable= Math Value Added (VA)									
	Current Year (Actual Results)								Previous year (Placebo)	
Ind. Var =	VA of Leavers		VA of Joiners		VA of Joiners - Leavers		Change in School Avg. VA		VA of Leavers	VA of Joiners
*Received **lower** grade at:*	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Bottom of the grade dist.										
C/D threshold	0.037	-0.159	1.232	2.008	1.053	1.795	0.021	0.042	0.058	0.059
	[0.3778]	[0.3605]	[0.4459]***	[0.9159]**	[0.4607]**	[0.9088]**	[0.0334]	[0.0374]	[0.3035]	[0.5071]
N	208	208	56	56	56	56	2,119	2,119	268	78
Top of the grade dist.										
B/C threshold	0.068	0.113	-1.496	-1.625	-1.588	-1.733	-0.016	-0.019	-0.120	0.251
	[0.1992]	[0.2114]	[0.6263]**	[0.6963]**	[0.6495]**	[0.7050]**	[0.019]	[0.0185]	[0.243]	[0.3795]
N	394	394	86	86	86	86	4,456	4,456	531	136
A/B threshold	-0.225	-0.146	-0.744	-0.998	-0.550	-0.765	-0.010	-0.019	-0.104	0.158
	[0.2275]	[0.2199]	[0.323]**	[0.3769]***	[0.3235]*	[0.3808]**	[0.0161]	[0.0168]	[0.2899]	[0.6418]
N	309	309	88	88	88	88	3,777	3,776	373	120
Dependent Var. Mean	-0.055	-0.058	-0.058	-0.056	-0.056	0.016	0.016	0.003	-0.067	0.010
Control for score*year* schooltype* accountability grade?	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
School covariates?		✓		✓		✓		✓		✓

Notes. Table presents regression discontinuity estimates of the effect of school accountability grades on the quality of the teachers that leave (leavers) and teachers that are hired by a given school (joiners). Specifically, for columns (1)-(6), (9), and (10): each observation is a teacher in a given year, the dependent variable is the math value added of the teachers in the sample, and the sample is the leavers from the school at the end of the year the school received a given accountability grade (columns (1) and (2)), the joiners to a school at the end of the year the school received a given accountability grade (columns (3)-(6)), the leavers from the school in the year before a school received a given accountability grade (column (9)- placebo), and the joiners to a school in the year before a school received a given accountability grade (column (10) - placebo). For columns (7)-(8), each observation is a teacher in a given year, and the dependent variable is the year to year change in average teacher quality at that teacher's school. Regressions use a bandwidth of 3 grade points. Standard errors are reported in brackets and clustered at the school level. School controls include controls for the average previous year's achievement; the percent of students that are black, hispanic, that receive free and reduced price lunch, and that are immigrants; 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 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 5. Regression Discontinuity Estimates of the Effect of School Accountability Grades on Joiner Experience and Education

Sample:	Joiners in current year (Actual Results)						Joiners in previous year (Placebo)		
Dependent Var.:	At least 2 Years Experience		At least 4 Years Experience		High Education (Has a Masters)		At least 2 Yrs Exp	At least 4 Yrs Exp	High Education
*Ind. Var. = Received **lower** grade at the:*	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Bottom of the grade distribution									
D/F or C/D thresholds (grouped)	0.015	0.044	0.009	0.031	-0.073	-0.058	-0.014	-0.020	0.024
	[0.07]	[0.07]	[0.06]	[0.06]	[0.06]	[0.05]	[0.05]	[0.05]	[0.04]
N	892	892	892	892	892	892	1,442	1,442	1,442
D/F Threshold	0.462	0.447	0.355	0.330	0.201	-0.041	0.040	0.084	0.180
	[0.08]***	[0.14]***	[0.09]***	[0.14]**	[0.10]*	[0.11]	[0.12]	[0.07]	[0.12]
N	235	235	235	235	235	235	451	451	451
C/D Threshold	-0.091	-0.072	-0.073	-0.090	-0.134	-0.128	-0.040	-0.105	0.013
	[0.08]	[0.08]	[0.07]	[0.08]	[0.07]*	[0.07]*	[0.07]	[0.06]*	[0.05]
N	675	675	675	675	675	675	1,027	1,027	1,027
Top of the grade distribution									
B/C or A/B Thresholds (grouped)	-0.041	-0.053	-0.052	-0.060	-0.042	-0.055	-0.052	-0.040	-0.024
	[0.05]	[0.04]	[0.04]	[0.04]	[0.03]	[0.03]*	[0.04]	[0.03]	[0.03]
N	2,342	2,342	2,342	2,342	2,342	2,342	3,425	3,425	3,425
B/C Threshold	-0.009	-0.008	-0.034	-0.033	0.002	-0.004	-0.075	-0.026	-0.041
	[0.07]	[0.06]	[0.06]	[0.06]	[0.05]	[0.05]	[0.05]	[0.04]	[0.04]
N	1,268	1,268	1,268	1,268	1,268	1,268	1,846	1,846	1,846
A/B Threshold	-0.108	-0.114	-0.091	-0.104	-0.106	-0.108	-0.036	-0.076	-0.019
	[0.07]	[0.07]*	[0.07]	[0.07]	[0.05]**	[0.05]**	[0.06]	[0.04]*	[0.04]
N	1,074	1,074	1,074	1,074	1,074	1,074	1,579	1,579	1,579
Dependent Variable Mean	0.37	0.37	0.24	0.24	0.20	0.20	0.27	0.18	0.16
Control for score*year*schooltype* accountability grade?	✓	✓	✓	✓	✓	✓	✓	✓	✓
School and Teacher covariates?		✓		✓		✓	✓	✓	✓

Notes. Table presents regression discontinuity estimates of the effect of school accountability grades on the characteristics of the teachers that are hired by a given school (joiners). Specifically, each observation is a teacher in a given year, and the dependent variable is that teacher's experience or education, and the sample is the joiners to a school at the end of the year the school received the accountability grade (columns (1)-(6)- actual results) or the joiners to a school in the year before a school received an accountability grade (columns (7)-(9) - placebo). Regressions use a bandwidth of 3 grade points. Standard errors are reported in brackets and clustered at the school level. School controls include controls for the average previous year's achievement; the percent of students that are black, hispanic, that receive free and reduced price lunch, and that are immigrants; 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 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 6. Regression Discontinuity Estimates of the Effect of School Accountability Grades on Achievement

Dependent Variable:	ELA Achievement		Math Achievement	
*Independent Var. = School received **lower** grade at the:*	(1)	(2)	(3)	(4)
Bottom of the grade distribution				
D/F or C/D thresholds (grouped)	0.001	0.031	0.022	0.022
	[0.08]	[0.06]	[0.08]	[0.05]
N	344	344	344	344
D/F Threshold	0.416	0.179	0.352	0.120
	[0.12]***	[0.09]**	[0.13]***	[0.10]
N	120	120	120	120
C/D Threshold	-0.131	-0.008	-0.051	0.049
	[0.08]*	[0.05]	[0.08]	[0.05]
N	283	283	283	283
Top of the grade distribution				
B/C or A/B Thresholds (grouped)	-0.037	-0.003	-0.008	0.032
	[0.05]	[0.03]	[0.05]	[0.03]
N	1,104	1,103	1,104	1,103
B/C Threshold	-0.006	0.015	0.017	0.037
	[0.07]	[0.04]	[0.07]	[0.04]
N	583	583	583	583
A/B Threshold	-0.100	-0.029	-0.090	0.002
	[0.08]	[0.04]	[0.07]	[0.04]
N	530	529	530	529
Control for schooltype*accountability grade* score?	✓	✓	✓	✓
School covariates?		✓		✓

Notes. The table presents regression discontinuity estimates of the effect of school accountability grades on student achievement. Regressions use a bandwidth of 6 grade points. The sample is all schools receiving accountability grades during the 2008-2009 and 2009-2010 school years, and each observation represents a school*year average. Standard errors are reported in brackets and clustered at the school level. The dependent variable is the average school-level mathematics or ELA test scores (standardized by the mean and standard deviation across all students taking the test in that year and grade) from the end of the schoolyear when the school received the accountability grade. Regressions are weighted by the number of students at each school that took the test. School controls include controls for the average previous year's achievement; the percent of students that are black, hispanic, that receive free and reduced price lunch, and that are immigrants; 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, 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%.

Table 7. Heterogeneity in the Regression Discontinuity Turnover Estimates by Teacher Destination

Dependent Variable:	1{Left}	1{Retired}	1{Transferred}	1{Left NYCDOE Classrooms}
*Independent Var. = School received **lower** grade at the:*	(1)	(2)	(3)	(4)
Bottom of the grade distribution				
D/F or C/D thresholds (grouped)	-0.037	-0.003	-0.006	-0.028
	[0.012]***	[0.004]	[0.008]	[0.008]***
N	17,932	17,979	17,932	17932
D/F Threshold	-0.041	-0.015	-0.014	-0.012
	[0.028]	[0.006]**	[0.018]	[0.018]
N	5,392	5,403	5,392	5,392
C/D Threshold	-0.026	0.003	-0.009	-0.020
	[0.012]**	[0.004]	[0.007]	[0.008]**
N	15,275	15,315	15,275	15,275
Top of the grade distribution				
B/C or A/B Thresholds (grouped)	0.002	0.001	-0.001	0.003
	[0.006]	[0.002]	[0.004]	[0.005]
N	63,968	64,091	63,968	63,968
B/C Threshold	0.003	0.001	-0.001	0.002
	[0.008]	[0.002]	[0.006]	[0.006]
N	34,386	34,455	34,386	34,386
A/B Threshold	0.004	0.001	-0.002	0.006
	[0.009]	[0.002]	[0.006]	[0.007]
N	30,029	30,083	30,029	30,029
Dependent variable mean	0.107	0.008	0.034	0.065
What proportionate coefficient would be for the grouped C/D and D/F thresholds	-0.037	-0.003	-0.012	-0.022
Control for score*year*schooltype*accountability grade?	✓	✓	✓	✓
School covariates?	✓	✓	✓	✓
Teacher covariates?	✓	✓	✓	✓

Notes. Table presents regression discontinuity estimates of the effect of school accountability grades on teacher turnover. The dependent variable for column (1) is an indicator for whether a teacher left their school in the summer after the accountability grade was received; columns (2) - (4) break up departures between the three ways teachers can leave the school (retirements, transfers, and stopping working in NYCDOE classrooms). 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. Regressions use a bandwidth of 6 grade points. Standard errors are reported in brackets and clustered at the school level. School controls include controls for the average previous year's achievement; the percent of students that are black, hispanic, that receive free and reduced price lunch, and that are immigrants; 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, and teacher gender. 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 8. Correlation Between Improvements in School Performance and Labor Market Outcomes

	Turnover Results			Joiner Results		
Dependent Variable	*Teacher Left School*			*Math Value-Added*		
Sample	*Incumbents*			*Joiners*		
Schools Near the:	(1)	(2)	(3)	(4)	(5)	(6)
Bottom of the grade distribution						
D/F or C/D thresholds (grouped)	-0.199	-0.082	-0.127	n/a	n/a	n/a
	[0.023]***	[0.063]	[0.064]**			
N	17,932	17,932	17,932			
D/F Threshold	-0.199	-0.166	-0.342	n/a	n/a	n/a
	[0.047]***	[0.143]	[0.122]***			
N	5,392	5,392	5,392			
C/D Threshold	-0.181	-0.084	-0.115	2.3	4.7	5.4
	[0.025]***	[0.073]	[0.075]	[0.73]***	[2.79]*	[2.73]**
N	15,275	15,275	15,275	57	56	56
Top of the grade distribution						
B/C or A/B Thresholds (grouped)	-0.101	-0.170	-0.173	0.0	1.4	1.6
	[0.012]***	[0.036]***	[0.035]***	[0.36]	[1.23]	[1.26]
N	63,970	63,968	63,968	174	174	174
B/C Threshold	-0.132	-0.215	-0.217	-0.5	4.0	4.7
	[0.015]***	[0.049]***	[0.049]***	[0.59]	[2.17]*	[2.73]*
N	34,386	34,386	34,386	86	86	86
A/B Threshold	-0.075	-0.111	-0.116	0.5	0.8	2.3
	[0.015]***	[0.051]**	[0.051]**	[0.47]	[1.58]	[1.69]
N	30,031	30,029	30,029	88	88	88
School covariates?		✓	✓		✓	✓
Teacher covariates?		✓	✓			
Accountability score control?			✓			✓

Notes. Table presents estimates of the correlation between improvements in school achievement and labor market outcomes, specifically, turnover at the end of the year (cols (1)-(3)) or the quality of joiners hired in the subsequent year (cols 4-6). Each row represents a separate regression using schools within a small bandwidth of a grade threshold (the bandwidth used is the same as the relevant bandwidth for the RD specifications—6 points for cols (1)-(3), 3 points for cols (4)-(6)). The coefficient presented is the coefficient on the average (between math and ela) school achievement of the school, measured in standard deviations of the student achievement distribution for the school year. For columns (1)-(3), 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, and each observation represents one teacher in a given year. For columns (4)-(6), the dependent variable is a teacher's math value-added; the sample is all joiners to a school at the end of the year the school received a given accountability grade, and each observation represents one teacher in a given year. Standard errors are reported in brackets and clustered at the school level. School controls include controls for the average previous year's achievement; the percent of students that are black, hispanic, that receive free and reduced price lunch, and that are immigrants; 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, and teacher gender. All data from 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

Panel A

	Dependent Variable=1{Teacher Left School}									
	Local Linear									
Bandwidth:	Base * 50% (3 points)		Base - 1 (5 points)		Base (6 points)		Base + 1 (7 points)		Base * 200% (12 points)	
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)	-0.014	-0.028	-0.024	-0.031	-0.027	-0.037	-0.019	-0.030	-0.018	-0.023
	[0.019]	[0.017]*	[0.015]	[0.013]**	[0.015]*	[0.012]***	[0.014]	[0.012]**	[0.013]	[0.011]**
N	7,798	7,798	16,314	16,314	17,932	17,932	19,924	19,924	31,484	31,484
P-Value: Spec Test	0.58	0.29	0.58	0.29	0.58	0.29	0.58	0.29	0.58	0.29
D/F Threshold	-0.056	-0.016	-0.059	-0.011	-0.078	-0.041	-0.069	-0.024	-0.060	-0.023
	[0.026]**	[0.033]	[0.023]**	[0.026]	[0.026]***	[0.028]	[0.023]***	[0.026]	[0.021]***	[0.019]
N	2,478	2,478	4,767	4,767	5,392	5,392	6,431	6,431	8,192	8,192
P-Value: Spec Test	0.91	0.34	0.91	0.34	0.91	0.34	0.91	0.34	0.91	0.34
C/D Threshold	0.001	-0.008	-0.002	-0.022	-0.009	-0.026	0.003	-0.019	-0.011	-0.019
	[0.022]	[0.021]	[0.015]	[0.013]*	[0.015]	[0.012]**	[0.014]	[0.012]*	[0.013]	[0.010]*
N	5,471	5,471	13,507	13,507	15,275	15,275	17,799	17,799	29,597	29,597
P-Value: Spec Test	0.54	0.15	0.54	0.15	0.54	0.15	0.54	0.15	0.54	0.15
Top of the grade distribution										
B/C or A/B Thresholds	-0.001	-0.005	0.008	0.005	0.006	0.002	0.009	0.005	0.007	0.000
	[0.011]	[0.011]	[0.008]	[0.007]	[0.007]	[0.006]	[0.007]	[0.006]	[0.006]	[0.005]
N	22,048	22,046	54,575	54,573	63,970	63,968	73,534	73,532	92,682	92,662
P-Value: Spec Test	0.53	0.80	0.53	0.80	0.53	0.80	0.53	0.80	0.53	0.80
B/C Threshold	-0.012	-0.008	0.002	0.004	0.002	0.003	0.007	0.006	0.010	0.006
	[0.016]	[0.013]	[0.011]	[0.009]	[0.010]	[0.008]	[0.009]	[0.007]	[0.007]	[0.006]
N	12,043	12,043	29,028	29,028	34,386	34,386	39527	39527	62837	62837
P-Value: Spec Test	0.22	0.29	0.22	0.29	0.22	0.29	0.22	0.29	0.22	0.29
A/B Threshold	0.005	-0.007	0.019	0.008	0.015	0.004	0.015	0.006	0.004	0.001
	[0.016]	[0.016]	[0.011]*	[0.010]	[0.010]	[0.009]	[0.010]	[0.009]	[0.007]	[0.006]
N	10,005	10,003	25,547	25,545	30,031	30,029	35672	35670	60732	60712
P-Value: Spec Test	0.34	0.82	0.34	0.82	0.34	0.82	0.34	0.82	0.34	0.82
Control for score*schooltype* year*accountability grade?	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
School and teacher covariates?		✓		✓		✓		✓		✓
Control for schooltype*(components of score)*year?										

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. The control function is linear for columns (1)-(10), quadratic for columns (11) and (12), and cubic for columns (13) and (14). Columns (15) and (16) control linearly for the underlying components of the accountability score. School controls include controls for the average previous year's achievement; the percent of students that are black, hispanic, that receive free and reduced price lunch, and that are immigrants; 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, and teacher gender. 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 9. Robustness of the Regression Discontinuity Turnover Estimates

Panel B

	Dependent Variable=1{Teacher Left School}					
	Quadratic		Cubic		Detailed Score Control	
Bandwidth:	12		12		6	
Ind. Var.=Received **lower** grade at the:	(11)	(12)	(13)	(14)	(15)	(16)
Bottom of the grade distribution						
D/F or C/D (grouped)	-0.020	-0.035	-0.020	-0.040	-0.012	-0.010
	[0.016]	[0.015]**	[0.021]	[0.018]**	[0.009]	[0.008]
N	31,484	31,484	31,484	31,484	17,932	17,932
P-Value: Spec Test	0.75	0.08	0.10	0.01	0.32	0.21
D/F Threshold	-0.032	-0.008	-0.037	-0.038	-0.028	-0.025
	[0.026]	[0.025]	[0.025]	[0.025]	[0.022]	[0.022]
N	8,192	8,192	8,192	8,192	5,392	5,392
P-Value: Spec Test	0.39	0.08	0.53	0.08	0.13	0.10
C/D Threshold	0.001	-0.019	0.004	-0.028	-0.012	-0.026
	[0.017]	[0.015]	[0.022]	[0.019]	[0.012]	[0.011]**
N	29,597	29,597	29,597	29,597	15,275	15,275
P-Value: Spec Test	0.22	0.07	0.28	0.07	0.10	0.30
Top of the grade distribution						
B/C or A/B Thresholds	0.006	0.001	0.002	0.000	0.004	0.003
	[0.008]	[0.007]	[0.011]	[0.010]	[0.004]	[0.003]
N	92,682	92,662	92,682	92,662	63,970	63,968
P-Value: Spec Test	0.51	0.97	0.24	0.88	0.67	0.47
B/C Threshold	0.000	0.005	0.002	0.004	-0.005	-0.002
	[0.011]	[0.009]	[0.014]	[0.011]	[0.009]	[0.008]
N	62837	62837	62837	62837	34386	34386
P-Value: Spec Test	0.06	0.46	0.06	0.28	0.57	0.41
A/B Threshold	0.010	0.002	0.019	0.005	-0.005	-0.003
	[0.011]	[0.010]	[0.014]	[0.013]	[0.008]	[0.007]
N	60,732	60,712	60,732	60,712	30,031	30,029
P-Value: Spec Test	0.59	0.84	0.58	0.86	0.15	0.48
Control for score*schooltype* year*accountability grade?	✓	✓	✓	✓	✓	✓ ✓✓
School covariates?		✓		✓		✓ ✓
Control for schooltype*(components of score)*year?					P	P

Table 10. Robustness of the Regression Discontinuity Joiner Quality Estimates

Panel A

	Math Value Added									
	Local Linear									
Bandwidth:	Base * 50% (1.5 points)		Base - 1 (2 points)		Base (3 points)		Base + 1 (4 points)		Base * 200% (6 points)	
*Ind. Var = Received **lower** grade at:*	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Bottom of the grade dist.										
C/D threshold	2.475	n/a	1.988	3.567	1.232	2.008	0.841	1.222	0.622	0.750
	[0.99]**		[0.91]**	[3.49]	[0.45]***	[0.92]**	[0.46]*	[0.77]	[0.40]	[0.55]
N	35		43	43	56	56	68	68	109	109
P-Value: Spec Test	0.65		0.65	0.62	0.65	0.62	0.65	0.62	0.65	0.62
Top of the grade dist.										
B/C threshold	-2.788	-6.720	-2.464	-3.235	-1.496	-1.625	-1.043	-1.090	-0.686	-0.895
	[1.55]*	[1.41]***	[1.18]**	[1.21]***	[0.63]**	[0.70]**	[0.45]**	[0.49]**	[0.35]*	[0.36]**
N	40	40	61	61	86	86	122	122	193	193
P-Value: Spec Test	0.85	0.66	0.85	0.66	0.85	0.66	0.85	0.66	0.85	0.66
A/B threshold	-0.309	0.110	-0.087	-0.515	-0.744	-0.998	-0.491	-0.682	-0.384	-0.339
	[0.59]	[1.42]	[0.52]	[0.73]	[0.32]**	[0.38]***	[0.29]*	[0.30]**	[0.26]	[0.27]
N	43	43	53	53	88	88	119	119	192	192
P-Value: Spec Test	0.86	0.68	0.86	0.68	0.86	0.68	0.86	0.68	0.86	0.68
Control for score*schooltype* year*accountability grade?	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
School covariates?		✓		✓		✓		✓		✓
Control for schooltype*(components of score)*year?										

Notes. Table presents regression discontinuity estimates. Each observation is a teacher in a given year, the dependent variable is the math value added of the teachers in the sample, and the sample is the joiners to a school at the end of the year the school received the accountability grade. The control function is linear for columns (1)-(10) and (17)-(20), quadratic for columns (11) and (12), and cubic for columns (13) and (14). Columns (15) and (16) control linearly for the underlying components of the accountability score. Standard errors are reported in brackets and clustered at the school level. School controls include controls for the average previous year's achievement; the percent of students that are black, hispanic, that receive free and reduced price lunch, and that are immigrants; 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, with the report card grades used being the report card that was received by the school during fall of the school year. 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

Panel B	Math Value Added						1{Above Median Math Value Added}		EB Value Added	
	Quadratic		Cubic		Detailed Score Control		Local Linear		Local Linear	
Bandwidth:	6		6		3		3		3	
*Ind. Var = Received **lower** grade at:*	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
Bottom of the grade dist.										
C/D threshold	0.289	0.283	2.108	1.002	0.859	1.355	1.055	1.348	1.063	1.689
	[0.53]	[0.56]	[0.84]**	[1.42]	[0.87]	[0.98]	[0.31]***	[0.49]***	[0.74]	[1.03]
N	109	109	109	109	56	56	56	56	56	56
P-Value: Spec Test	0.84	0.57	0.74	0.83	0.95	0.95	0.34	0.19	0.05	0.04
Top of the grade dist.										
B/C threshold	-1.759	-2.078	-2.049	-2.808	-1.868	-1.417	-0.543	-0.412	-0.499	-0.470
	[0.62]***	[0.60]***	[1.10]*	[1.05]***	[0.69]***	[0.81]*	[0.23]**	[0.31]	[0.48]	[0.56]
N	193	193	193	193	86	86	86	86	86	86
P-Value: Spec Test	0.37	0.15	0.29	0.17	0.57	0.23	0.80	0.75	0.55	0.42
A/B threshold	-0.489	-0.702	-0.612	-0.842	-0.718	-0.410	-0.358	-0.356	-0.552	-0.520
	[0.31]	[0.37]*	[0.42]	[0.52]	[0.33]**	[0.39]	[0.23]	[0.24]	[0.19]***	[0.27]*
N	192	192	192	192	88	88	88	88	88	88
P-Value: Spec Test	0.83	0.77	0.74	0.64	0.96	0.97	0.59	0.51	0.95	0.96
Control for score*schooltype* year*accountability grade?	✓	✓	✓	✓			✓	✓	✓	✓
School covariates?		✓		✓		✓		✓		✓
Control for schooltype*(components of score)*year?					✓	✓				

Appendix Table 1. Other Robustness Checks for the Regression Discontinuity Turnover Estimates

	Dependent Variable: *1*{Teacher Left School}							
Independent Var. =	Sample includes outliers		Include mid-year departures		2008 Only		2009 Only	
*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 (grouped)	-0.020	-0.033	-0.025	-0.036	-0.035	-0.037	-0.011	-0.038
	[0.018]	[0.015]**	[0.015]	[0.012]***	[0.018]*	[0.014]***	[0.023]	[0.023]
N	18,474	18,474	17,748	17,748	12,529	12,529	5,403	5,403
D/F Threshold	-0.060	-0.007	-0.078	-0.041	-0.077	-0.024	-0.080	-0.033
	[0.046]	[0.043]	[0.026]***	[0.028]	[0.038]**	[0.039]	[0.024]***	[0.027]
N	5,810	5,810	5,327	5,327	3,909	3,909	1,483	1,483
C/D Threshold	-0.018	-0.040	-0.004	-0.023	-0.019	-0.035	0.014	-0.003
	[0.016]	[0.014]***	[0.015]	[0.012]*	[0.018]	[0.013]***	[0.025]	[0.023]
N	15,610	15,610	15,123	15,123	10,466	10,466	4,809	4,809
Top of the grade distribution								
B/C or A/B Thresholds (grouped)	0.005	0.002	0.005	0.001	0.011	0.004	0.001	0.000
	[0.007]	[0.006]	[0.007]	[0.006]	[0.010]	[0.008]	[0.010]	[0.009]
N	63,970	63,968	63,470	63,468	33,525	33,525	30,445	30,443
B/C Threshold	0.002	0.003	0.000	0.001	0.001	-0.001	0.002	0.004
	[0.010]	[0.008]	[0.010]	[0.008]	[0.013]	[0.010]	[0.015]	[0.014]
N	34,386	34,386	34,105	34,105	20,138	20,138	14,248	14,248
A/B Threshold	0.015	0.004	0.016	0.004	0.034	0.019	0.002	-0.006
	[0.010]	[0.009]	[0.010]	[0.009]	[0.016]**	[0.014]	[0.013]	[0.012]
N	30,031	30,029	29,808	29,806	13,834	13,834	16,197	16,195
Control for schooltype*accountability grade* score?	✓	✓	✓	✓	✓	✓	✓	✓
School covariates?		✓		✓		✓		✓
Teacher covariates?		✓		✓				

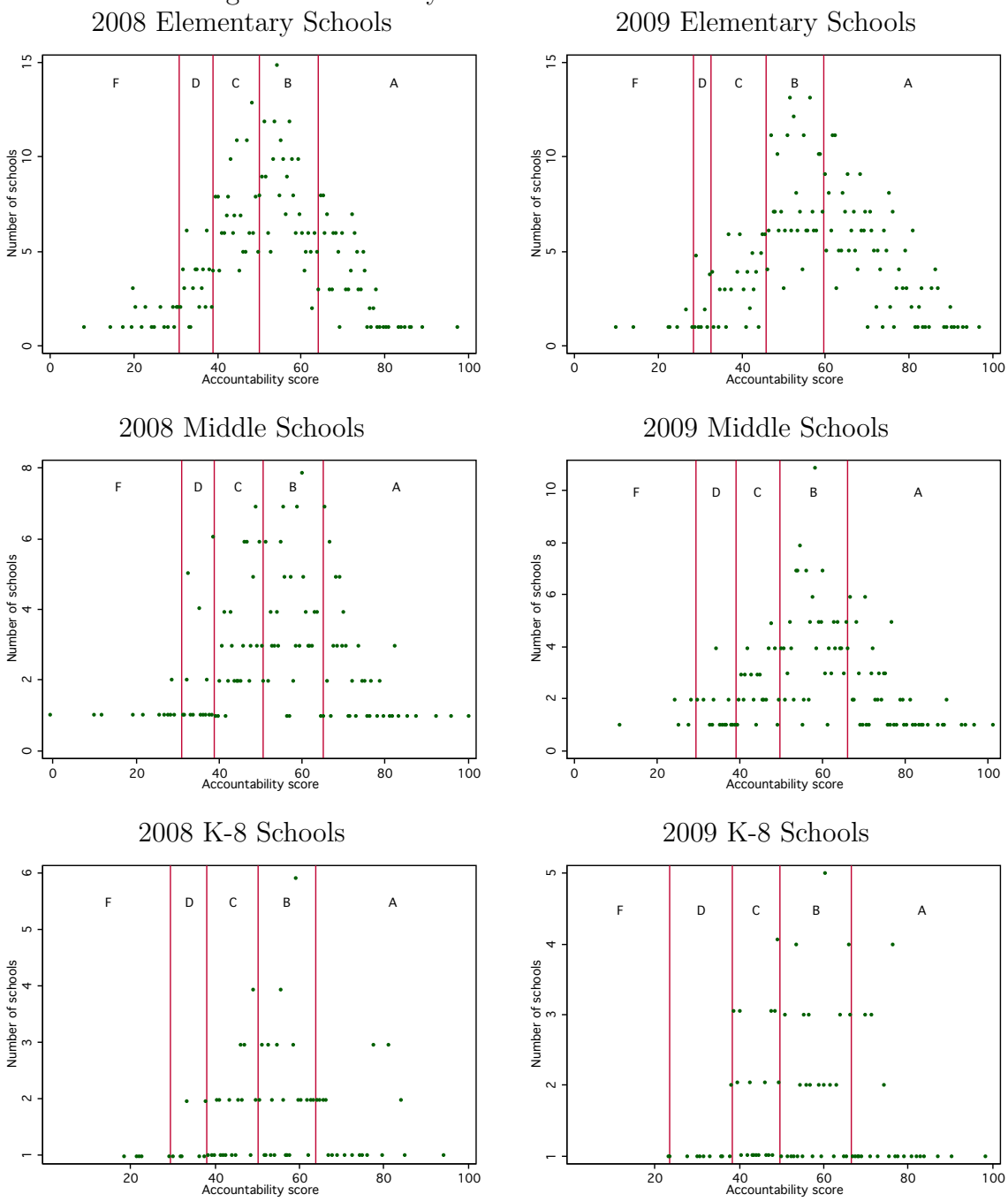
Notes. The table presents regression discontinuity estimates of the effect of school accountability grades on turnover. Regressions use a bandwidth of 6 grade points. Standard errors are reported in brackets and clustered at the school level. Each observation is a teacher in a given year, and the sample is all schools receiving accountability grades (columns (1) and (2)) or all schools that received accountability grades but do not appear to be undergoing restructuring (columns (3)-(8)). The dependent variable is an indicator that the teacher left the school, either between May of one year and November of the following year (columns (1) and (2), (5)-(8)) or between November and November (columns (3) and (4)). Columns (5) and (6) only include schools from the 2007-08 school year, and Columns (7) and (8) only include schools from the 2008-09 school year. School controls include controls for the average previous year's achievement; the percent of students that are black, hispanic, that receive free and reduced price lunch, and that are immigrants; 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, 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 2. Effect of School Accountability Grades on School Size

Dependent Variable:	Percent Change in Staff Size		Percent Change in Enrollment	
*Independent Var.= School received **lower** grade at the:*	(1)	(2)	(3)	(4)
Bottom of the grade distribution				
D/F or C/D thresholds (grouped)	0.02	0.01	-0.02	-0.03
	[0.020]	[0.019]	[0.032]	[0.030]
N	335	335	332	332
D/F Threshold	0.05	0.01	-0.08	-0.05
	[0.039]	[0.039]	[0.085]	[0.069]
N	114	114	111	111
C/D Threshold	0.01	0.00	0.04	0.03
	[0.020]	[0.019]	[0.027]	[0.027]
N	277	277	275	275
Top of the grade distribution				
B/C or A/B Thresholds (grouped)	-0.01	-0.01	-0.03	-0.03
	[0.009]	[0.008]	[0.016]*	[0.016]*
N	1,104	1,103	1103	1102
B/C Threshold	-0.01	-0.01	-0.04	-0.03
	[0.012]	[0.012]	[0.020]*	[0.020]*
N	583	583	582	582
A/B Threshold	-0.01	-0.01	-0.02	-0.02
	[0.013]	[0.013]	[0.025]	[0.024]
N	530	529	530	529
Control for schooltype*accountability grade* score?	✓	✓	✓	✓
School covariates?		✓		✓

Notes. The table presents regression discontinuity estimates of the effect of school accountability grades on school size. Regressions use a bandwidth of 6 grade points. Standard errors are reported in brackets and clustered at the school level. Each observation is a school in a given year. The dependent variable is the percent change in staff size (number of teachers) or enrolled students between the year that the school received the accountability grade and the following year (where 1.00 corresponds to 1 percentage point change). School controls include controls for the average previous year's achievement; the percent of students that are black, hispanic, that receive free and reduced price lunch, and that are immigrants; 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, 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%.

Figure A.1: Density of Schools Near Grade Thresholds



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.

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:

$$A_{ijgst} = \alpha + \beta_1 A_{i,j-1,g-1,s-1,t-1} + \beta_2 \bar{A}_{-i,j-1,g-1,t-1} + \beta_3 X_i + \tau_j + \tau_t + \tau_g + \tau_s + \eta_{jt} + \varepsilon_{ijgt} \quad (4)$$

where A_{ijgst} 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 school s and year t; $A_{i,j-1,g-1,s-1,t-1}$ 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 teachers, the year, the grade, and the school 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.[49] After estimation of equation (4), I standardize the τ_j terms and use this as my measure of teacher quality. I only estimate equation (4) using data from years before the institution of accountability in order to isolate teacher quality from teacher responses to accountability.

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 (4), 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.

I also construct empirical Bayes (EB) estimates of teacher value-added to check robustness. To do this, I follow the approach outlined in Kane and Staiger (2008) and Jackson (2009). The approach is described in the Online Appendix.

[49] Note that, despite the use of school fixed effects, I should be able to compare teacher fixed effects across different schools because there are many movers in the data. Moreover, for the RD analysis, schools on the border should have similar school fixed effects and so comparison of the fixed effects of teachers at the different schools will identify the effect of interest.

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 (2005) 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. Depending on the grade threshold and whether I used covariates, the optimal bandwidths according to this procedure ranged from 3-10 points when using the sample of all teachers and analyzing the turnover decision. Thus, I choose an intermediate (median) value as the base bandwidth for these regressions (6). When looking at the value-added outcomes and using as my samples either the joiners or the leavers, the optimal bandwidths ranged from 1-7, with most either 2 or 3, and I again use the median (3).

To check robustness, I also calculate a version of the Imbens-Kalyanarman (IK) optimal bandwidth (Imbens and Kalyanaraman, 2012).[50] For the turnover outcomes, these range from 2-6 with a median of 4, and for the value-added, they range from 1-5 with a median of 2; 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.

[50]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.

Appendix for Online Publication

Empirical Bayes Value-Added Estimates

I also construct empirical Bayes (EB) estimates of teacher value-added to check robustness. Although the estimates obtained by estimating equation 4 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 4, which is also the posterior mean with normally distributed errors.

I follow the approach outlined in Kane and Staiger (2008) and Jackson (2009). Consider the error term in equation 4, $w_{ijgt} \equiv \tau_j + \eta_{jt} + \varepsilon_{ijgt}$. 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 4 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, $\hat{\sigma_\tau^2} = Cov(\bar{w}_{jgt}, \bar{w}_{jgt-1})$.[51] 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: $\hat{\tau_j} = \sum_{j=1}^{J_j} \bar{w}_{jgt} \frac{(\sigma_\eta^2 + \sigma_\varepsilon^2/N_j)^{-1}}{\sum_{j=1}^{J_j}(\sigma_\eta^2 + \sigma_\varepsilon^2/N_j)^{-1}}$, where N_j is the number of students in classroom j and J_j is the number of classrooms that teacher j teaches.

Finally, I weight this estimate by an estimate of the precision of the teacher's effect to form the empirical Bayes estimate: $\hat{\tau_j}^{EB} = \hat{\tau_j} \frac{\sigma_\tau^2}{\sigma_\tau^2 + [\sum_{j=1}^{J_j}(\sigma_\eta^2 + \sigma_\varepsilon^2/N_j)^{-1}]^{-1}}$.

[51] 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.